

FACULTÉ DE MÉDECINE DE PARIS

ANNÉE 19 66 (1)

THÈSE

N° **176**

POUR LE

DOCTORAT EN MÉDECINE

(Diplôme d'État)

PAR

Edith LEDINOT - DUBREIL

Place réservée aux nom, prénoms, date et lieu de naissance
du Candidat

Présentée et soutenue publiquement le

1966

Les liaisons entre l'ou en

Place réservée au titre de la Thèse

Génétique Humaine
(Problèmes Théoriques)

Président : M. Hannu LAMY , Professeur

FACULTÉ DE MÉDECINE DE PARIS

ANNÉE 1966 (1)

THÈSE

N° — (2)

POUR LE

DOCTORAT EN MÉDECINE

(Diplôme d'État)

PAR

Edith LEDINOT-DUBREIL

Place réservée aux nom, prénoms, date et lieu de naissance
du Candidat

née le 22 Septembre 1936 à Paris XVI

Présentée et soutenue publiquement le 1966

Les liaisons entre loi en

Place réservée au titre de la Thèse

Génétique Humaine
(Problèmes Théoriques)

Président : M. Hannu LAMY, Professeur

FACULTE DE MEDECINE DE PARIS

ANNEE 1966

N°

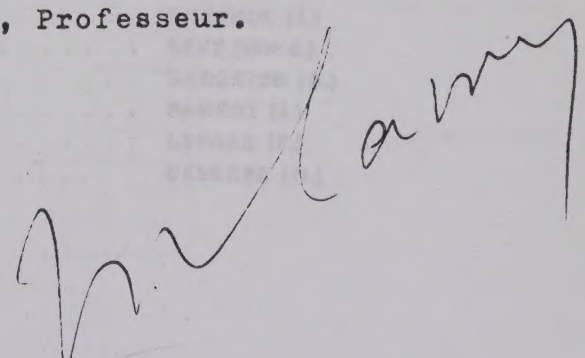
THESE
pour le
DOCTORAT EN MEDECINE
(Diplôme d'Etat)
par

Edith LEDINOT-DUBREIL
Née le 22-9-1936 à Paris XVI^e
Ancienne Externe des Hôpitaux de Paris

Présentée et soutenue publiquement le 1966.

Les liaisons entre loci en Génétique Humaine
(Problèmes théoriques)

Président : M. Maurice LAMY, Professeur.



PROFESSEURS HONORAIRES :

MM. Alajouanine (T.), Aubin (A.), Bénard (H.), Binet (L.), Bulliard (H.), Cathala (J.), Chabrol (E.), Chailley-Bert (P.), Debré (R.), Fey (B.), Galliard (H.), Gaudart d'Allaines (F. de), Gennes (L. de), Halphen (E.), Hazard (R.), Heuyer (G.), Laroche (G.), Lavier (G.), Lelong (M.), Lévy-Solal (E.), Lian (C.), Marchal (G.), Mathieu (P.), Monod (R.), Moreau (R.), Moulouquet (P.), Pasteur Vallery-Radot, Petit-Dutaillis (D.), Plédellèvre (R.), Quenu (J.), Renard (G.), Richet (C.), Senèque (J.), Strohl (A.), Tanon (L.), Verne (J.).

DOYEN Georges BROUET.

1^{er} ASSESSEUR Daniel BARGETON
2^e ASSESSEUR Jean GOSSET

I. — PROFESSEURS**1^o CHAIRES DE SCIENCES FONDAMENTALES****a) Professeurs titulaires :****MM.**

Anatomie	DELMAS (A.)
Anatomie et cytologie pathologiques	DELARUE (J.)
Anesthésiologie	BAUMANN (J.)
Bactériologie et virologie	TOURNIER (P.)
Biochimie médicale	JAYLE (M.-F.)
—	POLONOVSKI (J.)
Biologie appliquée à l'Education physique et aux Sports	MALMEJAC (J.)
Biologie médicale	FAUVERT (R.)
Biophysique médicale	BUGNARD (L.)
—	DOGNON (A.)
—	DJOURNO (A.)
Cancérologie médicale et sociale	DENOIX (P.)
Chimie pathologique	SCHAPIRA (G.)
Embryologie	TUCHMANN (H.)
Génétique fondamentale	LEJEUNE (J.)
Histoire de la médecine et de la chirurgie	COURY (Ch.)
Histologie	GIROUD (A.)
Hygiène et médecine préventive	BOYER (J.)
Médecine du Travail	DESOILLE (H.)
Médecine légale	DEROBERT (L.)
Microbiologie (Bactériologie et virologie)	FASQUELLE (R.)
Parasitologie	BRUMPT (L.-C.)
Pathologie chirurgicale	LAURENCE (G.)
—	KUSS (R.)
—	VERNE (J.-M.)
Pathologie exotique	N...
Pathologie expérimentale et comparée	MERKLEN (F.-P.)
Pathologie médicale	LAROCHE (Ol.)
—	BRICAIRE (H.)
—	PÉQUIGNOT (H.)
Pathologie respiratoire	TURIAF (J.)
Pathologie et thérapeutique générales	N...
Pharmacologie	CHEYMOL (J.)
—	LEVY (Mlle J.)
Physiologie	BARGETON (D.)
—	PARROT (J.)
Physiologie et pathologie obstétricales	LEPAGE (F.)
Radiologie médicale	DESGREZ (H.)

Technique chirurgicale et Chirurgie expérimentale
Thérapeutique
Thérapeutique appliquée

MM.
CAUCHOIX (J.)
CONTE (M.)
LAMOTTE (M.)

b) Professeurs à titre personnel :

Anatomie anthropologique
Anatomie pathologique
Bactériologie

Biochimie médicale

Biologie médicale

Histologie

Parasitologie

Pathologie expérimentale

Physiologie

Physique médicale

Radiologie médicale

OLIVIER (G.)
NEZELOF (Ch.)
BARBIER (P.)
DESGREZ (P.)
DREYFUS (J.-Cl.)
BOURLIÈRE (F.)
COUJARD (R.)
LARIVIÈRE (M.)
LOEPER (J.)
RICHET (G.)
DEJOURS (P.)
ROUCAYROL (J.-Cl.)
SCHERRER (J.)
KELLERSHOHN (L.)
GOUGEROT (L.)
TUBIANA (M.)
FISCHGOLD (H.)

2° CHAIRES DE SCIENCES CLINIQUES

a) Professeurs titulaires :

Clinique carcinologique chirurgicale (Necker)

Clinique cardiologique (Broussais)

— — (Boucicaut)

Cliniques chirurgicales
Beaujon
—
Cochin
Hôtel-Dieu
Saint-Antoine
Salpêtrière
Tenon

Clinique chirurgicale orthopédique et traumatologique (Cochin)

(Foch)

Clinique de chirurgie infantile et orthopédique (Enfants-Malades)

Clinique de chirurgie cardio-vasculaire (Broussais)

Clinique de chirurgie pleuro-pulmonaire (Laennec)

Clinique dermatologique (Saint-Louis)

Clinique endocrinologique (Pitié)

Clinique de Génétique médicale (Enfants-Malades)

Clinique gynécologique (Broca)

Clinique des maladies cutanées et syphilitiques (Saint-Louis)

Clinique des maladies des enfants (Enfants-Malades)

Clinique des maladies infectieuses (Claude-Bernard)

Clinique des maladies mentales et de l'encéphale (Sainte-Anne)

Clinique des maladies métaboliques (Necker)

Clinique des maladies du sang (Broussais)

Clinique des maladies du système nerveux (Salpêtrière)

Cliniques médicales
Bicêtre
Broussais
Cochin
Hôtel-Dieu
Lariboisière
Pitié
Saint-Antoine
Saint-Antoine

MM.

REDON (H.)
SOULIÉ (P.)
LENÈGRE (J.)
LORTAT-JACOB (J.-L.)
N...
LEGER (L.)
PATEL (J.)
GOSSET (J.)
SICARD (A.)
OLIVIER (Cl.)
MERLE d'AUBIGNÉ (R.)
PADOVANI (P.)
FÈVRE (M.)
DUBOST (Ch.)
MATHEY (J.)
DUPERRAT (F.)
DECOURT (J.)
LAMY (M.)
HUGUIER (J.)
DEGOS (R.)
TURPIN (R.)
MOLLARET (P.)
DELAY (J.)
HAMBURGER (J.)
BERNARD (J.)
CASTAIGNE (P.)
DEPARIS (M.)
MILLIEZ (P.)
BROUET (G.)
BARIÉTY (M.)
SÈZE (S. de)
DREYFUS (G.)
KOURILSKY (R.)
BOUDIN (G.)

Clinique d'hydrologie et de climatologie (Bichat)	DEBRAY (Ch.)
Clinique médicale infantile et pédiatrie sociale (Enfants-Malades)	MARIE (J.)
Clinique médicale propédeutique (Broussais)	JUSTIN-BESANÇON (L.)
Clinique médico-sociale du diabète sucré et des maladies de la nutrition (Hôtel-Dieu)	DEROT (M.)
Clinique de neuro-chirurgie (Pitié)	DAVID (M.)
Clinique neurologique (Salpêtrière)	GARCIN (R.)
Cliniques obstétricales	Baudelocque Saint-Antoine
Clinique obstétricale et gynécologique (Port-Royal)	LACOMME (M.)
Clinique ophtalmologique (Hôtel-Dieu)	MERGER (R.)
— — (Cochin)	VARANGOT (J.)
Clinique oto-rhino-laryngologique (Lariboisière)	OFFRET (G.)
Clinique de pédiatrie	BREGEAT (P.)
Clinique de pédiatrie et de puériculture (Saint-Vincent-de-Paul)	AUBRY (M.)
Clinique de pneumo-phtisiologie (Laennec)	MOZZICONACCI (P.)
Clinique de neuro-psychiatrie infantile (Salpêtrière)	THIEFFRY (S.)
Clinique de rhumatologie médicale et sociale (Cochin)	BERNARD (E.)
Clinique stomatologique (Salpêtrière)	MICHAUX (L.)
Clinique thérapeutique chirurgicale (Vaugirard)	GOSTE (F.)
Clinique thérapeutique médicale (Saint-Antoine)	DECHAUME (M.)
Clinique toxicologique	ROUX (M.)
Clinique traumatologique et orthopédique (R.-Poincaré)	LEMAIRE (A.)
Clinique urologique (Cochin)	GAULTIER (M.)
— — (Necker)	JUDET (R.)
Hygiène et clinique de la première enfance (Trousseau)	ABOULKER (P.)
Séméiologie médicale (A. Chantin)	COUVELAIRE (R.)
— — (Lariboisière)	LAPLANE (R.)
— — (Pitié)	DOMART (A.)
— — (Rothschild)	BOUVRAIN (Y.)
Séméiologie et clinique chirurgicales (Broussais)	RAMBERT (P.)
	WOLFROMM (R.)
	POILLEUX (F.)

b) Professeurs à titre personnel :

Hématologie	ANDRÉ (R.)
Hydrologie et climatologie	BOUSSER (J.)
Maladies infectieuses	GIBERTON (A.)
	BASTIN (R.)
	AZERAD (E.)
	BOUR (H.)
	CABROL (Ch.)
	GROSSIORD (A.)
	LAMBLING (A.)
Médecine	LAUNAY (Cl.)
	PAYET (M.)
	PERRAULT (M.)
	SIGUIER (F.)
	WORMS (R.)
Neuro-chirurgie	LE BEAU (J.)
Neurologie	LHERMITTE F.
Neuro-psychiatrie	PICHOT (P.)
Oto-Rhino-Laryngologie	MADURO (R.)
Pédiatrie	POLONOVSKI (Cl.)
	ROYER (P.)
Pneumo-Phtisiologie	KREIS (B.)
	MEYER (A.)
Rhumatologie	DELBARRE (F.)

II. — PROFESSEURS SANS CHAIRE

MM.		MM.	
Anatomie pathologique . . .	BOUYGOU (Ch.)	Hématologie	BILSKI-PASQUIER (G.)
Anesthésiologie	VOURC'H (G.)	Maladies infectieuses . . .	DUPONT (V.)
Bactériologie	CHRISTOL (D.)	Physique médicale	{ COURSAGET (J.) GREMY (F.)
Biochimie médicale	{ CARTIER (P.) GONNARD (P.)	Physiologie	GIRARD (Fr.)
Biologie médicale	HARTMANN (L.)	Stomatologie	CHAPUT (M ^{me} A.)
		Thérapeutique	CHASSAGNE (P.)

III. — PROFESSEURS DÉTACHÉS DANS LES UNIVERSITÉS ÉTRANGÈRES

Bactériologie	HAUDUROY (P.) Pr. s. ch
Pathologie chirurgicale	RUDLER (J.) Pr. s. ch.

IV. — PROFESSEUR ASSOCIÉ

Statistique médicale et épidémiologique . . SCHWARTZ (D.)

Secrétaire général de la Faculté	M. LENA (J.)
Conseillers administratifs	Mlle CHAINTREUIL (G.) M. FAYE (R.)
Conservateur en Chef de la Bibliothèque . .	{ M. le Docteur HAHN (A.)
Conservateur du Musée d'Histoire de la Médecine	
Conservateurs	{ Mlles DUMAITRE (P.), LAURENT (H.), LE NOIR (G.),
Bibliothécaires	{ Mmes CONTET (J.), FORGET (F.), Mlles DUROZOY (M.-F.), ROBERGE (D.), Dr SCHILLER (J.)

A mes Parents,

Ils sont à l'origine de mon orientation vers la Génétique.
Leur exemple et leur aide affectueuse m'ont encouragée
tout au long de mes études. Qu'ils trouvent ici un bien
faible témoignage de mon affection, de mon admiration,
et de ma profonde reconnaissance.

A Philippe , Emmanuel et Véronique ,
avec toute ma tendresse.

A Monsieur le Professeur Maurice LAMY ,

qui, dès mes premières années d'études médicales ,
a su me donner le goût de la Génétique. Je lui dois
ma spécialisation et , bien plus encore , toute une
formation médicale , intellectuelle et humaine ,
qu'il a guidée avec une inépuisable bienveillance .

Que cette thèse soit pour lui le modeste témoignage
de ma profonde gratitude .

A Monsieur le Professeur Agrégé Jean FREZAL

qui m'a proposé le sujet de cette thèse , et , tout au long de son exécution , n'a cessé de me prodiguer ses précieux conseils . Son aide et ses encouragements dans les moments difficiles m'ont permis de mener la tâche jusqu'au but fixé.

Qu'il trouve ici mes très sincères remerciements et l'expression de toute ma reconnaissance.

A mes Maîtres d'Externat dans les Hôpitaux de Paris :

Monsieur le Professeur DEPARIS ,
dont l'inlassable bonté et le merveilleux sens
clinique ont été pour moi le plus bel exemple.

Monsieur le Professeur Paul MILLIEZ ,
son enseignement incomparable et ses grandes
qualités humaines m'ont profondément marqué.
Ses suggestions et ses encouragements m'ont
bien souvent aidée.

Qu'il en soit ici très vivement remercié.

Monsieur le Professeur Raymond TURPIN et
Monsieur le Professeur Agrégé LAFOURCADE ,
qui , au cours d'une année d'Externat , m'ont
amenée à choisir avec un réel enthousiasme
de me spécialiser en Génétique Humaine .

A mes Maitres en Génétique de la Faculté des Sciences de Paris ;
je leur dois ma formation de chercheur ; qu'ils veuillent bien
trouver ici l'expression de ma reconnaissance et de mon profond
respect .

INTRODUCTION

- Historique.

En 1910, T.H.MORGAN observe que certains caractères n'obéissent pas à la seconde loi de MENDEL sur la ségrégation indépendante, mais, qu'au contraire, ils restent liés entre eux, et sont transmis ensemble, d'où le nom de "liaison" donné à ce phénomène.

L'analyse de ces faits devait apporter des arguments décisifs à la théorie chromosomique de l'hérédité. Il est apparu que les gènes sont répartis en autant de groupes qu'il existe de paires chromosomiques, et que, dans chaque groupe, les gènes sont alignés les uns à la suite des autres dans un ordre déterminé. Chez la Drosophile, et dans d'autres espèces animales ou végétales, ainsi que chez des organismes inférieurs où ces études ont été particulièrement poussées, il a été possible d'établir des cartes factorielles superposables aux cartes chromosomiques.

L'étude théorique du phénomène de liaison chez l'Homme, a été esquissée en 1926, par L.S.SNYDER. La première méthode statistique fut élaborée par F.BERNSTEIN, dont les travaux ont été repris et développés par L.HOGBEN et J.B.S.HALDANE, (1934), puis par D.J.FINNEY, dans une série de mémoires (1940, 1941, 1942). En 1935, R.A.FISHER a proposé la méthode des "scores efficaces", transformée par C.A.B.SMITH en méthode du rapport de probabilités. Plus récemment, N.E.MORTON (1955) a utilisé les principes de l'analyse séquentielle pour la recherche des liaisons chez l'Homme par le test des "lod-scores" ou scores logarithmiques.

- Définition et problèmes généraux.

Si les aspects théoriques de la question ont donné lieu à de multiples travaux, l'étude pratique des liaisons entre loci est à peine ébauchée chez l'Homme, et l'on n'en connaît que quelques exemples certains. Ceci tient aux difficultés du sujet.

Définition.

Le "linkage", premier nom donné par les auteurs Anglais au phénomène que nous allons étudier, est la liaison entre deux ou plusieurs loci. ceci signifie que ces loci sont portés par le même chromosome, et qu'ils

ils r

subissent pas la ségrégation indépendante. En d'autres termes , ceci veut dire que les gènes qu'ils portent, passent ensemble d'une génération à l'autre, plus souvent que ne le voudrait le hasard. Par extension , on peut considérer que tous les loci d'un même chromosome sont liés entre eux , bien que deux loci A et B suffisamment éloignés l'un de l'autre se comportent comme s'ils étaient indépendants. En effet , chez des organismes tels que la Drosophile , il est toujours possible de trouver un locus intermédiaire C, qui soit lié , d'une part à A , et d'autre part à B , d'où l'on déduit que A est bien lié à B. En Génétique Humaine on doit , pour l'instant s'en tenir à la définition stricte.

Liaison entre loci et association entre caractères.

Si deux loci sont liés , ils passeront généralement ensemble d'une génération à l'autre , et les caractères déterminés par les gènes qui occupent ces loci se manifesteront simultanément. La liaison entre les loci provoque donc une association entre certains caractères chez les membres d'une même famille. Mais il faut ajouter à cette proposition quelques réserves essentielles :

- 1/- Dans une famille donnée , l'association habituellement constatée peut être rompue. Ainsi , si A et B sont deux caractères associés , on pourra trouver des sujets qui possèdent A mais non B , et inversement. Il se produit des recombinaisons , conséquence des crossing-over survenant lors de la méiose , en nombre impair , entre les deux loci étudiés.
- 2/- Si l'on considère plusieurs familles , on constatera que ce ne sont pas toujours les mêmes caractères qui sont associés. Dans l'une , A sera associé à B , dans une autre il sera associé à b , allèle de B , etc ... Ceci s'explique par le fait que chaque locus porte l'un quelconque d'une série de gènes allèles.

En conséquence , si l'on étudie une population d'individus et non une série de familles , on ne retrouve plus aucune association entre les caractères. La fréquence des associations entre A d'une part , et B , b , ou b' d'autre part , est exclusivement fonction des fréquences respectives de ces différents gènes , et indépendante du fait que leurs loci sont situés sur un même chromosome. La démonstration mathématique de ce fait est donnée en annexe de ce chapitre. Toutefois , ceci n'est vrai que si un nombre suffisant de générations s'est écoulé pour que la population ait atteint son état d'équilibre. Celui-ci est atteint d'autant plus vite que

les loci en question sont moins proches l'un de l'autre. En effet, s'ils sont très rapprochés, la probabilité qu'il se produise un crossing-over entre eux est extrêmement faible. Les recombinaisons seront exceptionnelles et les loci et les gènes qu'ils portent resteront associés pour de très nombreuses générations, soit en phase, soit en répulsion.

La liaison entre les loci est donc une forme très particulière de l'association entre les caractères, et la seule constatation d'une telle coïncidence ne suffit pas pour évoquer la proximité, sur un même chromosome, des gènes responsables. Beaucoup d'autres possibilités existent, qu'il convient de rappeler brièvement :

a) - Deux caractères sont associés s'ils relèvent de la même cause, par exemple, lorsqu'ils sont dus à un gène à effet pléiotrope. Cette association sera constante si la manifestation du gène est régulière. Dans le cas contraire, des dissociations peuvent être observées, qu'il ne faut pas confondre avec des recombinaisons.

b) - Le facteur responsable du premier caractère n'agit qu'en présence du second, ou, du moins, son action peut être facilitée par sa présence. C'est le cas de l'association au sexe, qui doit être soigneusement distinguée de la liaison au sexe. Ainsi, et pour des raisons différentes, on n'observera tel ou tel caractère déterminé par un gène autosomique que dans un sexe, l'autre étant épargné, mais transmetteur. Il en est ainsi de la goutte, de l'hémochromatose et de la calvitie, qui atteignent principalement les hommes.

c) - L'association peut résulter d'une interaction entre les facteurs responsables. Ainsi, l'antigène Lewis et le caractère Secréteur, sont déterminés par des gènes non allèles, dont les loci ne sont pas liés. Cependant, la manifestation du premier est influencée par le second: les sujets sécréteurs fabriquent peu d'antigène Le(a), et ont des globules rouges Le (a-).

d) - L'association entre les caractères n'implique pas nécessairement un rapport de causalité entre les phénomènes observés. Elle peut être due à l'hétérogénéité des populations étudiées ou à une erreur systématique dans le recensement des échantillons, (phénomène de stratification de L.S.PENROSE).

e) - Certaines associations sont encore mal comprises, comme celles qui ont été constatées entre certaines maladies et des groupes sanguins, qui sont souvent confondues avec une liaison entre loci, avec la quelle

elles n'ont pourtant aucun rapport. Ici , l'association se retrouve aussi bien dans la population que dans les familles. Elle porte toujours sur les mêmes caractères , par exemple le groupe O et l'ulcère du duodénum.

Pour prouver l'existence d'une liaison entre deux loci , il faut démontrer que, dans une famille , ou dans une série de familles , les caractères déterminés par les gènes portés par ces loci , ne subissent pas la ségrégation indépendante. Dire qu'il y a ségrégation indépendante signifie que , dans la descendance d'un couple , le premier caractère sera aussi souvent associé au second que séparé de lui ; ou encore , que le pourcentage de recombinaisons entre ces caractères est égal à 50%. La valeur du pourcentage de recombinaisons , comprise entre 0 et 50% , donne une estimation de l'étroitesse de la liaison, puisque sa valeur varie dans le même sens que la distance qui sépare les loci.

Les données du problème.

L'étude des liaisons entre loci se présente différemment selon que l'on considère les autosomes ou les chromosomes sexuels. Pour ce qui est des caractères autosomiques , l'étude des ségrégations établit (ou écarte) l'existence d'une liaison. Elle ne permet en aucune façon de situer ces loci sur un chromosome particulier. En revanche , cette localisation pourrait être déduite , dans une certaine mesure et dans certains cas bien particuliers , de l'étude cytologique et biochimique des aberrations chromosomiques. Il s'agit là d'une voie de recherche ouverte récemment et encore peu explorée. Au contraire , il est relativement aisé de démontrer qu'un gène est porté par le chromosome X. Le problème est alors de situer son emplacement par rapport à d'autres loci , c'est-à-dire de dresser une carte chromosomique. Ce problème n'est facile qu'en apparence. Au demeurant , il est probable que cet ordre n'est pas absolument constant, mais qu'il est , au contraire , différent chez certains sujets par suite de remaniements chromosomiques , d'inversions par exemple . Dans certains cas d'autre part , il est difficile de distinguer une hérédité liée au sexe d'un phénomène de limitation au sexe . Ainsi , on ne peut distinguer , à l'examen d'une généalogie , une tare léthale récessive liée au sexe, d'une tare léthale dominante autosomique qui ne s'exprimerait que chez les garçons (testicule féminisant par exemple). La démonstration de l'existence d'une liaison entre le locus de ce gène, et un locus marqueur porté

par le chromosome X permettrait de résoudre élégamment le problème.

Bien entendu , on ne peut observer de recombinaisons entre deux caractères liés au sexe , que dans la descendance masculine des femmes , qui , seules , possèdent deux chromosomes X . On a cru pendant longtemps que le chromosome X et le chromosome Y possédaient une partie homologue au niveau de laquelle pourraient s'effectuer des échanges , comme on l'observe dans certaines espèces animales et végétales . J.B.S.HALDANE a étudié , d'un point de vue théorique , la transmission et la ségrégation de caractères contrôlés par des gènes supposés situés sur cette partie commune , caractères dits partiellement liés au sexe , et il a montré que certains traits observés dans les familles humaines répondaient à ces conditions . Cependant d'autres interprétations sont possibles , et , de fait , les observations cytologiques vont à l'encontre de cette hypothèse , puisque les chromosomes X et Y ne s'apparient pas lors de la méiose . On admet donc généralement que les chromosomes sexuels n'ont pas de partie paire. N.E.MORTON , par l'analyse statistique , a du reste pu éliminer l'hypothèse d'une hérédité "partiellement liée au sexe" dans les cas précédemment évoqués par HALDANE . Il s'en suit que les caractères déterminés par des gènes portés par le chromosome Y auront un mode de transmission très particulier : ils n'apparaîtront que chez l'homme qui , seul , possède un chromosome Y , et seront transmis à toute sa descendance mâle . Toutefois , si l'on connaît le rôle important du chromosome Y dans la différenciation sexuelle masculine , on ne connaît que peu d'exemples de caractères monofactoriels répondant à ce mode de transmission dit holandrique .

En pratique , on cherche à mettre en évidence une liaison entre le locus d'un gène responsable d'une maladie quelconque , et le locus d'un gène marqueur , responsable d'une variation normale . Rien ne s'oppose , en théorie , à ce que l'on étudie la liaison entre des loci contrôlant deux maladies , sinon le fait que leur survenue simultanée est exceptionnelle . Les caractères étudiés doivent répondre à certaines conditions précises : a)- Le caractère principal doit être responsable d'une anomalie malformative ou biologique précise , et d'un diagnostic aisé . Le gène qui le contrôle doit avoir une manifestation régulière à l'état hétérozygote . On pourra donc étudier :

- les caractères dominants rares à manifestation régulière .
- parmi les caractères récessifs , ceux dont le gène responsable

est régulièrement décelable à l'état hétérozygote , comme le gène de la galactosémie .

b)- Les marqueurs , doivent répondre à des conditions voisines : déterminisme génétique simple , diagnostic facile , manifestation régulière . On ne peut donc utiliser des caractères physiques tels que la couleur des yeux , la pilosité digitale , etc ... , dont le déterminisme génétique est mal connu . La sensibilité à la thiourée n'est pas non plus un très bon caractère , des erreurs de classification étant possibles entre sujets goûteurs et non goûteurs . En pratique , seuls peuvent être retenus les groupes érythrocytaires , et certaines protéines ou enzymes érythrocytaires ou sériques . Tous n'ont pourtant pas la même valeur . Celle-ci dépend de l'abondance ou de la rareté des anti-sérums , c'est-à-dire de la possibilité d'effectuer la détermination . Elle dépend aussi de la fréquence des différents allèles dans la population . Si cette dernière est homogène , tous les sujets , ou presque , ont le même génotype , et le groupe considéré n'apporte aucune information . De ce point de vue , il est clair que la multiplicité des allèles , qui augmente le polymorphisme , est avantageuse . Il en est ainsi du facteur Rhésus , ou du groupe MNS , dont le déterminisme génétique peut être assimilé à des systèmes d'allèles multiples , dont chacun est formé d'une combinaison particulière de sous-groupes : $CDe=R_1$, $cDe=R_2$, $cDe=R_3$, $cde=r$...

Il serait préférable , au surplus , de pouvoir reconnaître les hétérozygotes ; mais ceci n'est pas possible directement pour tous les systèmes . Le génotype peut toutefois être déduit , dans un certain nombre de cas , de l'étude des ascendants et descendants . Par exemple , un sujet de groupe A dont le père , ou un enfant , est de groupe O , a nécessairement le génotype AO . Mais si l'on procède ainsi , il faut , sous peine d'erreur , ne pas tenir compte , dans le calcul du pourcentage de recombinaison , des enfants qui ont servi à la détermination du génotype de l'un de leurs parents , à moins d'introduire une correction (D.J.FINNEY) .

On voit , d'après ce qui précède , que le nombre des marqueurs utilisables est encore très restreint . De fait , on n'en compte guère plus d'une quinzaine , c'est-à-dire moins que le nombre de paires chromosomiques . Nous donnons un tableau de ces divers marqueurs à la page 35 .

c)- Généalogies .

Pour étudier les liaisons , il faut disposer de familles comprenant au moins deux générations , c'est-à-dire , au minimum un mariage et ses

descendants . Les familles étudiées sont sélectionnées en fonction de la présence du caractère principal et , dans le cas d'un gène dominant , il s'agira habituellement d'une famille issue de l'union d'une hétérozygote et d'un homozygote pour l'allèle récessif (ou l'inverse) , ce dernier contrôlant le caractère. Ce mariage est appelé un test-cross . Le mode de recensement des familles ne suit pas nécessairement une méthode statistique précise en ce qui concerne le locus principal , mais , dans ces familles , les marqueurs sont toujours distribués au hasard . Cependant , seuls certains types de mariages peuvent apporter une information. Le plus favorable d'entre eux est le double test-cross , c'est-à-dire le mariage entre un double hétérozygote pour le caractère principal : Pp , et pour le caractère marqueur : Mm , et un double homozygote récessif ppmm . Si M est dominant par rapport à m , la fréquence de ces unions est maxima quand les fréquences p de M et q de m sont dans le rapport 1/3 , c'est-à-dire $p=0,25$ et $q=0,75$. Dans ce cas , le plus favorable , 1/5 seulement des mariages retenus pour la ségrégation du caractère principal seront également des test-cross pour le caractère marqueur . Cette situation "idéale" n'est qu'approchée pour certains groupes . De plus , la dominance de l'un des caractères peut faire que ce type de mariage ne soit pas reconnu , si d'autres sujets de la famille ne permettent pas de distinguer Mm de MM comme il a été indiqué plus haut . La situation est meilleure quand on peut nettement distinguer l'hétérozygote , et utiliser , selon certaines méthodes statistiques , d'autres types de mariages tels que :

- test-cross x inter-cross
Pp.Mm x pp.Mm
- inter-cross x test-cross
Pp.Mm x Pp.mm
- inter-cross x inter-cross
Pp.Mm x Pp.Mm

Un mariage test-cross x test-cross : Pp.mm x pp.Mm ne donnerait , au contraire , aucune information .

Nous avons dit que l'étude des liaisons exige que l'on connaisse au moins deux générations de chaque famille étudiée ; mais il est clair que les généalogies étendues présentent de nombreux avantages . En particulier , le génotype de certains sujets peut être établi par référence au phénotype de ses ascendants ou collatéraux . Il est souvent difficile de savoir

si les gènes étudiés aux deux loci d'une femme hétérozygote double , sont en position de phase ou de répulsion . La connaissance des phénotypes des grand-parents de cette femme permet de résoudre le problème . Au demeurant , le nombre de familles nécessaire est moins élevé quand celles-ci sont de grande dimension . Enfin , ces familles étendues permettent d'aborder le problème de l'hétérogénéité . Dans une famille donnée , il est hautement probable que le caractère principal résulte de la même mutation , transmise d'une génération à l'autre . Si une famille donne isolément une information suffisante , il est possible de comparer les résultats obtenus à ceux observés dans d'autres familles . De cette façon , il sera éventuellement possible de distinguer des mutations qui ont le même effet phénotypique, si l'une d'elles est liée à un marqueur mais non l'autre . On sait ainsi qu'il existe deux types d'elliptocytose , le locus de l'un est lié au locus Rh, mais non celui de l'autre . Cependant toutes les données enregistrées sur une vaste généalogie n'apportent pas nécessairement une information utile . Dans la majorité des cas , on devra se contenter d'analyser une série de familles de dimensions réduites, qui ne donnent évidemment, a priori, aucune garantie d'homogénéité .

En dernier lieu , il convient d'attirer l'attention sur les difficultés d'ordre statistique et mathématique que soulève l'étude des liaisons entre loci, chez l'homme. La première a trait aux conditions de recensement des échantillons et à la sélection des familles , qui ne se fait pas selon les lois du hasard, comme elle le devrait , mais en fonction de la présence d'individus atteints . La probabilité de sélection d'une famille est alors proportionnelle au nombre de sujets atteints dans la famille . Ceci introduit un biais dans la distribution , et oblige à utiliser des corrections . Cependant , l'erreur ainsi introduite est faible si le caractère étudié n'est pas soumis à l'influence de la sélection , (R.A.FISHER , 1935) . L'utilisation du phénotype de certains enfants pour déterminer le génotype des parents , oblige , sous peine d'erreur , à éliminer ces enfants du calcul du "score", ou à employer des corrections laborieuses . Enfin , ainsi que le fait remarquer C.A.B.SMITH , la probabilité pour que deux loci tirés au hasard soient liés, est approximativement de $1/23$, c'est-à-dire voisine de $4/100$. S'il existait un test statistique parfait , permettant de mettre en évidence toutes les liaisons , ce test , utilisé à la limite de confiance habituelle de 95% ,

affirmerait à tort l'existence d'une liaison dans 5 cas sur 100 . Ce qui revient à dire que l'on annoncerait l'existence d'une liaison plus souvent à tort qu'à raison . Il faut donc choisir une limite de confiance plus sévère , avec un risque d'erreur de 1/1000 seulement par exemple .

ANNEXE A L'INTRODUCTION
EVOLUTION SIMULTANEE DE DEUX PAIRES DE GENES
DANS UNE POPULATION PANMICTIQUE

Considérons deux couples d'allèles , (A , a) d'une part , (B , b) d'autre part . Nous pouvons définir neuf génotypes et les classer dans le tableau ci-dessous :

$$(1) \quad \begin{bmatrix} \text{AABB} & \text{AABb} & \text{AAbb} \\ \text{AaBB} & \text{AaBb} & \text{Aabb} \\ \text{aaBB} & \text{aaBb} & \text{aabb} \end{bmatrix}$$

au quel nous associerons le tableau homologue de leurs fréquences relatives dans la population étudiée , supposée infinie :

$$(2) \quad F = \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{bmatrix} = [f_{ij}]$$

Nous avons alors :

$$\text{fréquence du génotype AA} = f_{11} + f_{12} + f_{13}$$

$$\text{fréquence du génotype Aa} = f_{21} + f_{22} + f_{23}$$

$$\text{fréquence du génotype aa} = f_{31} + f_{32} + f_{33}$$

$$\text{fréquence du génotype BB} = f_{11} + f_{21} + f_{31}$$

$$\text{fréquence du génotype Bb} = f_{12} + f_{22} + f_{32}$$

$$\text{fréquence du génotype bb} = f_{13} + f_{23} + f_{33}$$

$$\text{avec } \sum_i \sum_j f_{ij} = 1 \quad (i=1,2,3 ; j=1,2,3)$$

Nous associerons de même aux quatre types de gamètes possibles :

AB , Ab , aB , ab , les tableaux suivants :

$$(3) \quad \begin{bmatrix} \text{AB} & \text{Ab} \\ \text{aB} & \text{ab} \end{bmatrix}$$

$$(4) \quad \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix}$$

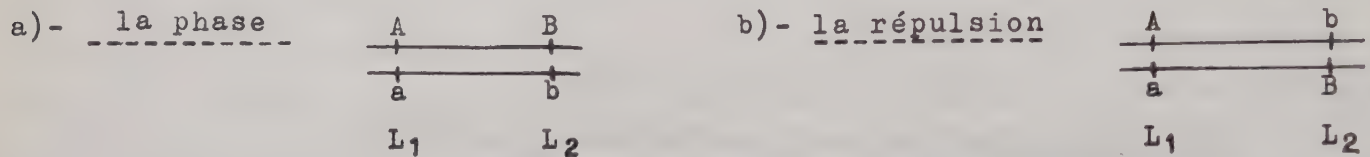
avec : $\sum_i \sum_j g_{ij} = 1$ ($i=1,2$; $j=1,2$)

Si nous désignons par p_1 , p_2 , q_1 , q_2 , les fréquences respectives des gènes A , a , B , b dans l'ensemble de la population , nous avons :

$$(5) \quad \begin{aligned} p_1 &= g_{11} + g_{12} & p_2 &= g_{21} + g_{22} \\ q_1 &= g_{11} + g_{21} & q_2 &= g_{12} + g_{22} \end{aligned}$$

- Variations des g_{ij} au cours des générations successives .

La liaison entre les deux loci L_1 et L_2 peut se présenter sous deux modes différents :



Nous dirons qu'il y a phase , lorsque A et B d'une part , a et b d'autre part , sont portés par le même chromosome .

Il y aura répulsion si A est associé à b sur l'un des chromosomes et a à B sur l'autre chromosome de la paire .

Si la probabilité de recombinaison entre les deux loci L_1 et L_2 est x , ($0 \leq x \leq 1$) , les parents hétérozygotes doubles AaBb produisent quatre types de gamètes : AB , Ab , aB , ab , dans des proportions indiquées par les expressions suivantes , obtenues par application des deux principes fondamentaux du Calcul des Probabilités :

-a)- cas de la phase .

fréquence des gamètes AB : $\frac{1-x}{2}$

fréquence des gamètes Ab : $\frac{x}{2}$

fréquence des gamètes aB : $\frac{x}{2}$

fréquence des gamètes ab : $\frac{1-x}{2}$

-b)- cas de la répulsion .

fréquence des gamètes AB : $\frac{x}{2}$

fréquence des gamètes Ab : $\frac{1-x}{2}$

fréquence des gamètes aB : $\frac{1-x}{2}$

fréquence des gamètes ab : $\frac{x}{2}$

Les parents qui possèdent au moins une paire de gènes à l'état homozygote , c'est-à-dire les sujets de génotypes AaBB , AABb , aaBb , Aabb , AaBB , aabb , aaBB , AAbb , produisent les mêmes gamètes , et dans les mêmes proportions , qu'il y ait liaison ou non . Il est par exemple évident que , dans tous les cas , (AaBB) $\longrightarrow$ $1/2$ (AB) + $1/2$ (aB)

Tableau de passage d'une génération, n , à la suivante $n+1$

Gamètes fournis par la génération n. ♀ ♂		AB g_{11}^n	Ab g_{12}^n	aB g_{21}^n	ab g_{22}^n
AB g_{11}^n	ZYGOTES génération $n+1$.	$\frac{AB}{AB}$	$\frac{Ab}{AB}$	$\frac{aB}{AB}$	$\frac{ab}{AB}$
	GAMETES génération $n+1$.	AB 1	AB $\frac{1}{2}$ Ab $\frac{1}{2}$	AB $\frac{1}{2}$ aB $\frac{1}{2}$	AB $\frac{1-x}{2}$ ab $\frac{1-x}{2}$ aB $\frac{x}{2}$ Ab $\frac{x}{2}$
Ab g_{12}^n	ZYGOTES génération $n+1$.	$\frac{AB}{Ab}$	$\frac{Ab}{Ab}$	$\frac{aB}{Ab}$	$\frac{ab}{Ab}$
	GAMETES génération $n+1$.	AB $\frac{1}{2}$ Ab $\frac{1}{2}$	Ab 1	AB $\frac{x}{2}$ ab $\frac{x}{2}$ Ab $\frac{1-x}{2}$ aB $\frac{1-x}{2}$	Ab $\frac{1}{2}$ ab $\frac{1}{2}$
aB g_{21}^n	ZYGOTES génération $n+1$.	$\frac{AB}{aB}$	$\frac{Ab}{aB}$	$\frac{aB}{aB}$	$\frac{ab}{aB}$
	GAMETES génération $n+1$.	AB $\frac{1}{2}$ aB $\frac{1}{2}$	AB $\frac{x}{2}$ ab $\frac{x}{2}$ Ab $\frac{1-x}{2}$ aB $\frac{1-x}{2}$	aB 1	aB $\frac{1}{2}$ ab $\frac{1}{2}$
ab g_{22}^n	ZYGOTES génération $n+1$.	$\frac{AB}{ab}$	$\frac{Ab}{ab}$	$\frac{aB}{ab}$	$\frac{ab}{ab}$
	GAMETES génération $n+1$.	AB $\frac{1-x}{2}$ ab $\frac{1-x}{2}$ Ab $\frac{x}{2}$ aB $\frac{x}{2}$	Ab $\frac{1}{2}$ ab $\frac{1}{2}$	aB $\frac{1}{2}$ ab $\frac{1}{2}$	ab 1

Ceci posé , considérons une population initiale , F_0 , dont la composition gamétique est définie par :

$$\begin{bmatrix} g_{11}^0 & g_{12}^0 \\ g_{21}^0 & g_{22}^0 \end{bmatrix}$$

et cherchons à déterminer la composition gamétique de la génération suivante , F_1 , représentée par :

$$\begin{bmatrix} g_{11}^1 & g_{12}^1 \\ g_{21}^1 & g_{22}^1 \end{bmatrix}$$

Calculons par exemple g_{11}^1 , fréquence des gamètes (AB) dans F_1 . Ces gamètes peuvent être formés par les sujets de génotypes (AABB) , (AABb) , (AaBB) , (AaBb) de F_0 . Les trois premiers sont homozygotes pour au moins une paire de gènes ; par conséquent , d'après la remarque qui précède :

$$\begin{aligned} \text{Pr (AABB)} &\rightarrow (\text{AB}) &= 1 \\ \text{Pr (AABb)} &\rightarrow (\text{AB}) &= 1/2 \\ \text{Pr (AaBB)} &\rightarrow (\text{AB}) &= 1/2 \end{aligned}$$

Pour le dernier génotype , il faut distinguer deux cas :

$$\begin{aligned} \text{a)- Pr (AaBb) en phase} &\rightarrow (\text{AB}) &= \frac{1-x}{2} \\ \text{b)- Pr (AaBb) en répulsion} &\rightarrow (\text{AB}) &= \frac{x}{2} \end{aligned}$$

D'autre part , par hypothèse , les gamètes s'unissent au hasard pour former les sujets de la génération F_1 . On a donc :

$$\begin{aligned} \text{Pr (AABB)} &= (g_{11}^0)^2 \\ \text{Pr (AABb)} &= 2g_{11}^0 g_{12}^0 \\ \text{Pr (AaBB)} &= 2g_{11}^0 g_{21}^0 \\ \text{Pr (AaBb) en phase} &= 2g_{11}^0 g_{22}^0 \\ \text{Pr (AaBb) en répulsion} &= 2g_{12}^0 g_{21}^0 \quad \text{etc ...} \end{aligned}$$

Les autres probabilités sont faciles à déduire , de la même façon , du tableau de la page précédente .

De ces relations , et par application des principes des probabilités totales et composées , il résulte que :

$$g_{11}^1 = (g_{11}^0)^2 + g_{11}^0 g_{12}^0 + g_{11}^0 g_{21}^0 + (1-x)g_{11}^0 g_{22}^0 + xg_{12}^0 g_{21}^0$$

ou encore :

$$\begin{aligned} g_{11}^1 &= g_{11}^0 (g_{11}^0 + g_{12}^0 + g_{21}^0 + g_{22}^0) - x (g_{11}^0 g_{22}^0 - g_{12}^0 g_{21}^0) \\ &= g_{11}^0 - x (g_{11}^0 g_{22}^0 - g_{12}^0 g_{21}^0) \end{aligned}$$

Posons :

$$(6) \quad d_0 = g_{11}^0 g_{22}^0 - g_{12}^0 g_{21}^0$$

d_0 est le déséquilibre entre les sujets en phase et ceux en répulsion à la génération F_0 . d_0 n'est autre que le déterminant de la matrice qui définit

la composition gamétique de F_0 . Avec cette notation nous obtenons :

$$(7) \quad g_{11}^1 = g_{11}^0 - x d_0$$

et par des calculs absolument analogues :

$$(8) \quad g_{12}^1 = g_{12}^0 + x d_0$$

$$(9) \quad g_{21}^1 = g_{21}^0 + x d_0$$

$$(10) \quad g_{22}^1 = g_{22}^0 - x d_0$$

On vérifie aisément que $g_{11}^0 + g_{12}^0 + g_{21}^0 + g_{22}^0 = g_{11}^1 + g_{12}^1 + g_{21}^1 + g_{22}^1 = 1$

- Conséquences .

-1/- Nous avons :

$$g_{11}^1 + g_{12}^1 = g_{11}^0 + g_{12}^0 = p_1$$

Donc , les fréquences des gènes , dans une population panmictique et infinie , non soumise à l'influence des mutations et de la sélection , sont invariables au cours des générations successives .

-2/- Posons :

$$d_1 = \begin{bmatrix} g_{11}^1 & g_{12}^1 \\ g_{21}^1 & g_{22}^1 \end{bmatrix}$$

d_1 peut aussi s'écrire :

$$d_1 = \begin{bmatrix} g_{11}^0 - x d_0 & g_{12}^0 + x d_0 \\ g_{21}^0 + x d_0 & g_{22}^0 - x d_0 \end{bmatrix} = (1-x) d_0$$

d'où , au cours des générations successives :

$$g_{11}^1 = g_{11}^0 - x d_0$$

$$g_{11}^2 = g_{11}^1 - x d_1 = g_{11}^0 - x(1-x) d_0$$

$$g_{11}^3 = g_{11}^2 - x d_2 = g_{11}^0 - x(1-x)^2 d_0$$

$$g_{11}^n = g_{11}^{n-1} - x d_{n-1} = g_{11}^{n-1} - x(1-x)^{n-1} d_0$$

Additionnons membre à membre , il vient :

$$g_{11}^n = g_{11}^0 - x d_0 \left[1 + (1-x) + (1-x)^2 + \dots + (1-x)^{n-1} \right]$$

soit , en calculant la somme de la progression géométrique entre crochets

$$(9) \quad g_{11}^n = g_{11}^0 - \left[1 - (1-x)^n \right] d_0$$

On obtiendrait de la même façon :

$$(10) \quad g_{12}^n = g_{12}^0 + \left[1 - (1-x)^n \right] d_0$$

$$(11) \quad g_{21}^n = g_{21}^0 + \left[1 - (1-x)^n \right] d_0$$

$$(12) \quad g_{22}^n = g_{22}^0 - \left[1 - (1-x)^n \right] d_0$$

Discussion :

.1/- $x=0$.

Il y a liaison absolue , tout se passe comme si l'on avait affaire à

une série multiallélique :

$$g_{11}^n = g_{11}^0 \quad ; \quad g_{12}^n = g_{12}^0 \quad ; \quad g_{21}^n = g_{21}^0 \quad ; \quad g_{22}^n = g_{22}^0$$

Les fréquences gamétiques , et par conséquent aussi les fréquences génotypiques , restent invariables au cours des générations successives .

.2)- $x=1$.

Le crossing - over entre L_1 et L_2 est une certitude .

$$g_{11}^n = g_{11}^0 - d_0 \quad \quad g_{12}^n = g_{12}^0 + d_0$$

$$g_{21}^n = g_{21}^0 + d_0 \quad \quad g_{22}^n = g_{22}^0 - d_0$$

si $d \neq 0$, l'équilibre est atteint en une génération .

si $d = 0$, les fréquences gamétiques gardent les valeurs qu'elles avaient en F_0 .

.3)- $0 < x < 1$.

Théoriquement , l'équilibre n'est jamais atteint . Cependant , comme $1-x < 1$, le terme $(1-x)^n$ tend vers zéro lorsque n tend vers l'infini , d'autant plus vite que x est plus proche de 1 , c'est-à-dire que les loci L_1 et L_2 sont plus éloignés l'un de l'autre sur le chromosome . Il en résulte que les fréquences gamétiques tendent vers les limites :

$$(1) \quad \hat{g}_{11} = g_{11}^0 - d_0 \quad \hat{g}_{12} = g_{12}^0 + d_0 \quad \hat{g}_{21} = g_{21}^0 + d_0 \quad \hat{g}_{22} = g_{22}^0 - d_0$$

La rapidité de la convergence dépend de la valeur de x .

Pour $x = 1/2$, cas de l'indépendance :

$$\Delta_n = [g_{ij}^n - \hat{g}_{ij}] = (1-x)^n [d_0] = \frac{[d_0]}{2^n}$$

comme $[d_0] \leq 1/4$, on a $\frac{[d_0]}{2^n} \leq \frac{1}{2^{n+2}}$, quantité inférieure à 10^{-3} pour $n < 5$.

Pour $x = 1/4$ et $n = 10$, Δ_n est de l'ordre de 10^{-2} .

- Conséquence pratique fondamentale :

La conclusion la plus importante à tirer de l'étude qui précède est que les fréquences gamétiques , lorsque la population a atteint l'état d'équilibre , $\hat{g}_{11}$, $\hat{g}_{12}$, $\hat{g}_{21}$, $\hat{g}_{22}$, sont indépendantes de l'existence ou non d'une liaison , et , dans le cas où il y a liaison , de son intensité .

Il en résulte que , dans une population panmictique , infinie , non soumise à l'influence des mutations et de la sélection , les fréquences zygotiques sont entièrement déterminées par les fréquences des gènes , que ces gènes soient localisés ou non sur le même chromosome . La liaison entre deux loci ne peut donc pas , nous l'avons vu , être décelée par l'examen des génotypes ou des phénotypes dans la population générale . Deux caractères , conditionnés par deux paires de gènes dont les loci sont liés , étant distribués , à l'équilibre , comme si les loci étaient indépendants , il est vain de vouloir , comme cela a souvent été tenté ,

déterminer les liaisons entre les loci étudiés par l'étude de la corrélation entre ces deux caractères .

Au cours des pages qui suivront , nous admettrons toujours implicitement ce qui n'est pas prouvé , que la population humaine a atteint son état d'équilibre .

CHAPITRE I : LES PROBLEMES THEORIQUES .

La recherche des liaisons entre loci soulève , en Génétique Humaine , un certain nombre de problèmes théoriques délicats que nous allons maintenant envisager successivement .

-1/- Qu'est-ce qu'un locus ?

Question à laquelle il est difficile de répondre sans donner au préalable une définition de la notion de gène . Nous laisserons les spécialistes de Génétique Physiologique et Biochimique étudier ce problème , incontestablement l'un des plus difficiles et des plus passionnants . Il n'est que d'ouvrir les traités de Génétique , pour entendre parler de gène de structure , de gène de contrôle , de gène modificateur , de cistron , d'opéron etc ...

Schématiquement nous entendrons par gène , une unité de fonction , c'est à dire , la quantité d'information génétique contrôlant la synthèse d'une chaîne polypeptidique .

Dans ces conditions , on peut dire que le locus correspondant est l'emplacement linéaire chromosomique occupé exactement par une unité de fonction ou par les allèles de celle-ci .

Ceci est évidemment assez simpliste , mais il serait inutile d'aller plus avant dans la structure fine du gène pour le but que nous nous proposons .

-2/- Combien y a-t-il de loci chez l'homme ?

Les résultats récents de la Génétique moléculaire nous permettent , sous certaines hypothèses , de faire une estimation approximative du nombre de loci existant dans l'espèce humaine (F.VOGEL , 1964) .

a)- La chaîne α de l'hémoglobine contient 141 acides aminés , la chaîne β en contient 146 , ce qui correspond à un poids moléculaire d'environ 17.000 pour chacune . Selon l'hypothèse du code de triplets , ceci signifie que les chaînes α et β sont déterminées respectivement par 423 et 438 paires de nucléotides . D'après la loi de SVEDBERG , de nombreuses protéines sont constituées de "sous-unités" du même ordre de grandeur (PM = 17.500 environ) . Il semblerait donc qu'un gène de structure α , en moyenne , une longueur égale à celle de 450 paires de nucléotides environ .

b)- Le poids de l'ensemble (haploïde) des chromosomes contenus dans un spermatozoïde humain est d'environ $2,72 \times 10^{-12}$ g ; dans les leucocytes , ce poids , correspondant cette fois à un ensemble diploïde , est de $6,23 \times 10^{-12}$ g ; il est de $5,84 \times 10^{-12}$ g dans les lymphocytes . D'autres études ont montré que la teneur en ADN est constante dans toutes les cellules au repos d'une espèce donnée , cellules qui possèdent le même nombre de chromosomes . En revanche , ce poids varie avec le degré de ploïdie . Il semble également vrai que la majeure partie de cet ADN forme le support de l'information génétique , même si , dans quelques cellules , de petites fractions peuvent avoir une autre fonction . Dans les calculs qui suivent, la quantité totale d'ADN contenue dans l'ensemble haploïde des chromosomes humains est estimée à environ 3×10^{-12} g .

c)- Habituellement , les variantes génétiques des hémoglobines humaines ne diffèrent que par un seul acide aminé . Un gène de structure contrôle la synthèse d'un seul type de polypeptide . On peut penser que ceci s'applique aussi bien à d'autres types de protéines déterminées génétiquement . Ceci veut dire aussi que l'information génétique fournie par ces gènes de structure ne peut être présente qu'une fois et une seule . Comme nous l'avons vu plus haut , le contenu en ADN des cellules diploïdes est à peu de choses près égal au double du contenu des cellules haploïdes (spermatozoïdes) . Nous supposons donc que l'information génétique totale est présente à un exemplaire et un seulement . Il est facile de calculer qu'une paire de nucléotides comportant une molécule d'adénine , une de thymine deux désoxyriboses et deux résidus phosphorés , a un poids moléculaire de $1,025 \times 10^{-21}$ g $\approx 1 \times 10^{-21}$ g . Ainsi , l'ensemble haploïde (3×10^{-12} g) , contient environ 3×10^9 paires de nucléotides . En admettant que , en moyenne , un gène de structure est composé de 450 nucléotides , nous arrivons à une estimation de $6,7 \times 10^6$ loci . Cette estimation paraît très élevée et il serait peut-être nécessaire de raisonner autrement . Dans les chromosomes géants de certains diptères , qui montrent un degré élevé de polyténie , on a pu montrer que les bandes observées sont des unités de fonction contrôlant chacune la synthèse d'une chaîne polypeptidique . A l'intérieur de ces bandes , une seule chromatide a une longueur moyenne de 50.000 paires de nucléotides selon W.BEERMANN . Cette estimation est environ 100 fois plus grande que celle donnée plus haut, et conduirait à 6×10^4 loci , approximativement , chez l'homme , ce qui est en bien meilleur accord avec les hypothèses faites jusqu'à présent . Mais il reste

La proportion de recombinaisons , x_0 , est directement observable par l'étude des mariages et de leurs descendance ; elle est égale à la somme des probabilités d'un nombre impair de crossing-over sur le segment considéré, soit :

$$x_0 = p_1 + p_3 + p_5 \dots$$

La distance sur la carte , x , peut être définie , nous le verrons au paragraphe suivant , comme l'espérance mathématique du nombre total de crossing-over sur le segment $L_1 L_2$, soit :

$$x = p_1 + 2p_2 + 3p_3 + 4p_4 + \dots$$

ce que l'on peut exprimer plus clairement en définissant une série , de terme général s_n , représentant la fréquence de n crossing-over ou plus, sur le segment $L_1 L_2$, de sorte que l'on a :

$$s_n = \sum_{i=n}^{\infty} p_i \quad p_n = s_n - s_{n+1}$$

Les équations :

$$x_0 = s_1 - s_2 + s_3 - s_4 + \dots$$

$$x = s_1 + s_2 + s_3 + s_4 + \dots$$

expriment , sous forme de séries , les deux variables génétiques qui nous intéressent ; des deux , seule x_0 est directement observable .

Toute théorie correcte du crossing-over doit fournir les moyens d'exprimer la série entière des valeurs de p ou de s en fonction de la position des marqueurs situés sur le chromosome envisagé , ainsi que de leurs positions respectives par rapport au centromère et aux extrémités du chromosome . Il s'agit donc d'exprimer x en fonction de x_0 . La première solution de ce problème , que nous ne donnerons pas ici dans le détail , a été apportée par J.B.S. HALDANE sous la forme :

$$x_0 = x_0 (x + dx) = x_0 (x) + dx - 2c_m (x_0) . x_0 (x) . dx$$

$c_m (x_0)$ est la "coïncidence marginale" de HALDANE .

Puis , il y a quelques années , une formule extrêmement efficace a été donnée par le mathématicien KOSAMBI :

$$2x_0 = \tan h(2x)$$

La formule de KOSAMBI conduit à un théorème d'additivité de la fraction de recombinaison x_{1+2} sur la somme de deux segments 1 et 2 dont les proportions de recombinaison sont respectivement x_1 et x_2 :

$$x_{1+2} = \frac{x_1 + x_2}{1 + 4x_1 x_2}$$

Cette formule peut être vérifiée expérimentalement dès lors que l'on dispose de trois marqueurs . L'approximation est très bonne, bien que l'on ne tienne pas compte de la position du segment considéré par rapport

au centromère ce qui , pourtant , influence beaucoup les phénomènes d'interférence dont nous parlerons au dernier paragraphe de ce chapitre .

Donc , même si cette formule est assez précise pour exprimer de façon satisfaisante la relation entre x et x_0 , elle ne peut en aucun cas être l'expression d'une théorie complète et correcte du crossing-over . En effet , si elle est suffisante pour établir les relations entre trois marqueurs , elle devient inefficace lorsque l'on considère quatre loci , ou davantage . Ainsi , pour quatre marqueurs par exemple , il existe huit couples de génotypes complémentaires , et la présence de sept constantes est nécessaire pour exprimer leurs fréquences relatives . Une formule comme la précédente ne peut donner qu'une approximation sur la fraction de recombinaisons entre trois loci . Seule la série complète de terme général s_n peut fournir une information théoriquement complète, sur les recombinaisons entre un nombre quelconque de loci . Pour K.MATHER , la formation des chiasmas sur une paire de chromosomes , au stade à quatre chromatides , est considérée comme un processus aléatoire partant du centromère, de telle façon que l'intervalle entre le centromère et le chiasma le plus proche de lui, a une distribution de fréquence bien définie . Il en est de même pour l'intervalle entre le premier chiasma et le second , et ainsi de suite . Il est évident qu'une fois ces distributions connues , il est possible de calculer les fréquences de toutes les configurations possibles , en spécifiant la probabilité d'apparition de toute suite de chiasmas sur la même chromatide . Il semble qu'un assez bon accord avec les observations qui ont été faites soit obtenu par l'introduction d'une métrique , qu'il ne faut pas confondre avec la distance sur la carte , ou avec la distance réelle . Grâce à cette métrique , l'interférence se manifeste de la même façon tout au long du chromosome . Nous étudierons ceci plus en détail à la fin de ce chapitre . Il semble suffisant , tout au moins pour avoir une bonne approximation , de considérer le centromère comme un point de chiasma obligatoire . Grâce à la métrique dont nous venons de parler , il devient possible de faire un calcul théorique de la série $s_1 , s_2 , s_3 , \dots s_n$, pour tout couple de loci donné , ainsi que des valeurs de x et x_0 , qui apportent toute l'information voulue sur la proportion de recombinaison dans une région quelconque d'un chromosome donné .

Reste maintenant à répondre à la question la plus importante : le crossing-over existe-t'il ? Il semble en effet qu'au moins un certain nombre

de recombinaisons soit dû , non pas à des crossing-over , mais à un autre phénomène désigné sous le nom de conversion . Nous laisserons les spécialistes répondre à cette question et , dans tout ce qui suivra , nous admettrons que les recombinaisons observées n'ont pas d'autre cause que le crossing-over .

-5/- Relations entre pourcentage de recombinaisons et "distance".

Le problème de la carte .

On convient d'utiliser le pourcentage de recombinaisons entre deux loci comme une mesure de leur distance . Cette distance , voisine de zéro si les loci sont très rapprochés , croît vers $1/2$ lorsqu'ils s'éloignent l'un de l'autre sur le chromosome . Théoriquement x varie entre 0 et 1 ; l'intervalle $1/2 - 1$ concerne le cas très particulier de l'interférence négative c'est-à-dire le cas , pratiquement exceptionnel , où la présence d'un crossing-over favorise la réalisation d'autres crossing-over dans son voisinage . En réalité on peut considérer que x varie seulement entre 0 et $1/2$, la valeur maxima , $1/2$, correspondant au cas de l'indépendance ou d'un éloignement tel sur le chromosome , que les deux loci se comportent comme s'ils étaient indépendants . Nous reparlerons de l'interférence un peu plus loin .

Le pourcentage de recombinaison est une assez bonne mesure directe de la distance pour des groupes de loci très proches les uns des autres , car il possède alors des propriétés d'additivité . Ainsi , pour trois loci A , B et C disposés dans cet ordre , le pourcentage des recombinaisons survenant entre A et C sera approximativement égal à la somme des pourcentages de recombinaisons entre A et B d'une part , entre B et C d'autre part . Cette règle d'additivité ne s'applique plus cependant , lorsque les loci ne sont plus proches l'un de l'autre . La densité des points d'échange ou chiasmas constitue alors une meilleure unité de mesure . En revanche , si nous définissons la "distance sur la carte" entre deux loci quelconques , comme l'espérance mathématique du nombre de points d'échange sur le segment considéré , nous aurons une quantité intrinsèquement additive , que les loci soient proches ou non . Il en est ainsi parce que l'espérance mathématique d'une somme de plusieurs variables est égale à la somme des espérances mathématiques de chacune de ces variables . De plus , ceci est encore vrai si ces variables ne sont pas statistiquement indépendantes . Une distance sur la carte ainsi définie resterait donc additive même en la présence d'interférence . Dans le cas le plus simple , où il n'y a pas

d'interférence , la relation entre le pourcentage de recombinaison et la distance sur la carte est extrêmement simple : supposons que l'ordre réel des trois loci sur le chromosome soit ABC . Soit x_1 le pourcentage de recombinaisons entre A et B , x_2 entre B et C ; nous allons montrer que x_{1+2} , pourcentage de recombinaisons entre A et C est donné par l'expression

$$x_{1+2} = x_1 + x_2 - 2x_1x_2 \quad (\text{ formule de TROW , 1913 })$$

En effet , $x_{1+2} = x_1(1-x_2) + x_2(1-x_1) = x_1 + x_2 - 2x_1x_2$

On voit aisément que si x_1 et x_2 sont tous deux petits , on peut , en première approximation , négliger le double produit $2x_1x_2$ et écrire :

$$x_{1+2} \approx x_1 + x_2$$

De nombreuses vérifications expérimentales de cette règle ont été faites , en particulier sur *Drosophila Melanogaster* .

Mais il n'y a , actuellement , aucune raison de penser que les chromosomes sont parfaitement homogènes en ce qui concerne leurs aptitudes au crossing-over . Nous ne pouvons pas admettre , a priori , qu'il existe nécessairement une étroite correspondance entre la distance sur la carte et la distance , réelle , sur le chromosome , entre deux loci quelconques . Les difficultés surgissent lorsque l'on essaie de tenir compte d'un certain degré d'interférence , c'est-à-dire du fait qu'un crossing-over inhibe la survenue d'autres crossing-over dans son voisinage . Les distances sur la carte restent additives , mais ne peuvent plus être calculées directement . Elles doivent être déduites , compte tenu de certaines hypothèses appropriées , des pourcentages de recombinaisons observées . Nous envisagerons un peu plus loin , assez rapidement , un certain nombre de méthodes , élémentaires et assez empiriques , utilisables dans le cas où il y a interférence , que celle-ci soit positive ou négative . Pour différentes raisons ces méthodes ne sont pas parfaitement satisfaisantes . Elles ne donnent pas , par exemple , d'information sur la possibilité de recombinaisons multiples sur de longs segments . On sait aussi que l'intensité des phénomènes d'interférence varie avec la distance au centromère . Il faut donc , si l'on veut tenir compte de cette réalité , utiliser une formule plus générale , qui utilise une métrique spéciale . Grâce à cette métrique , nous l'avons vu plus haut , l'action de l'interférence devient uniforme tout au long du chromosome . Ce procédé présente de nombreux avantages , théoriques et pratiques . Cette métrique possède naturellement la propriété d'additivité et permet le calcul des pourcentages de recombinaisons et des distances sur la carte pour toute paire de loci , à la seule

condition de connaître leur position par rapport au centromère . Ce modèle mathématique , plus général , permet aussi , sous réserve de quelques modifications , de tenir compte du fait que les bras chromosomiques sont de longueur finie . Quelques autres versions , légèrement différentes , de cette théorie mathématique , sont possibles ; mais , actuellement , il est encore impossible de décider lequel de ces modèles est le mieux adapté à l'usage courant .

Pour conserver la propriété d'additivité pour les distances moyennes et grandes , nous avons vu qu'il est nécessaire de travailler en termes d'espérance mathématique du nombre de crossing-over , plutôt qu'en pourcentage de recombinaisons observées , qui mesure seulement la fréquence d'un nombre impair de crossing-over . Le procédé mathématique dont nous venons de parler permet de situer deux loci quelconques l'un par rapport à l'autre mais ne donne aucun renseignement sur leur position par rapport au centromère . Ce site est pourtant d'une importance fondamentale , et , " bien que dépourvu de gènes , il peut sans doute être considéré comme un point d'échange obligatoire " , selon N.T.J.BAILEY .

Alors que la grande majorité des loci obéit au concept des " groupes de liaison " , indépendants les uns des autres , un petit nombre de liaisons anormales ont été observées . Dans certaines circonstances , des loci qui paraissent habituellement appartenir à des groupes de liaison différents , et , à l'intérieur de ces groupes , satisfont à l'hypothèse de l'arrangement linéaire des loci , peuvent parfois se comporter de la même façon que s'ils étaient réellement liés . Ce phénomène est appelé " quasi-liaison " . Il a été observé pour la première fois par W.H.GATES (1926) au cours d'études portant sur le croisement de deux espèces de souris , *Mus musculus* et *Mus bacterianus* . Ce phénomène pourrait être expliqué , (D.MICHIE , 1953) , par une tendance des centromères *musculus* à migrer vers l'un des pôles , alors que les centromères *bacterianus* migrent vers l'autre pôle , lors de la première division méiotique chez l'hybride . Il pourrait , en quelque sorte , y avoir un phénomène d'attraction entre les centromères de même origine .

En 1953 également , Mrs M.E.WALLACE a observé des quasi-liaisons à l'intérieur même de l'espèce *Mus musculus* . Sir R.A.FISHER a suggéré une explication analogue , reposant sur une orientation non au hasard des centromères , et a nommé ce processus " affinité " . Il est évident que

si , lors de la méiose , les centromères migrent vers les pôles, autrement qu'au hasard , les loci liés à ces centromères, se comporteront également de façon non aléatoire .

L'analyse des divers types de mariages conduit à une estimation de la fraction de recombinaison, et fournit une information directe sur la position du centromère par rapport aux autres loci situés sur le même chromosome . Lorsque , dans certaines circonstances , nous obtenons des résultats statistiques significativement en faveur d'une liaison entre deux loci qui , pourtant , pour d'autres raisons , semblent être situés sur des chromosomes différents , il peut s'agir d'un tel phénomène d'affinité centromérique .

- Le concept mathématique de la distance sur la carte .

La distance sur la carte doit être une fonction additive du pourcentage de recombinaison .

Nous pouvons écrire :

$$1 - 2x_{1+2} = (1 - 2x_1)(1 - 2x_2)$$

La forme générale de cette relation , pour r segments consécutifs , soit r+1 loci , est , de la même façon :

$$1 - 2x_{1+2+\dots+r} = \prod_{i=1}^r (1 - 2x_i)$$

Il devient alors évident, qu'en passant aux logarithmes, on obtient une fonction additive convenant parfaitement . Posons :

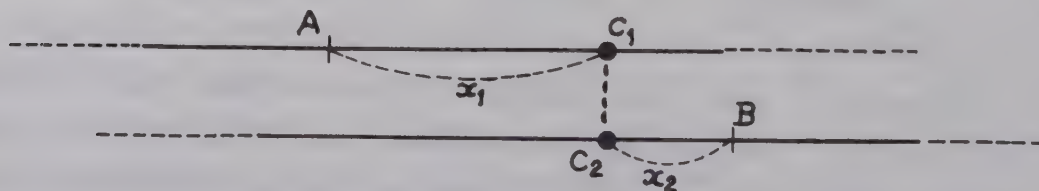
$$y(x) = k \ln(1 - 2x)$$

où k est une constante arbitraire . Il est particulièrement commode de choisir $k = -1/2$ de façon à ce que x et y(x) soient pratiquement équivalents pour des petites valeurs de x . Nous obtenons ainsi la " fonction carte " , y(x) , définie par :

$$y(x) = -1/2 \ln(1 - 2x)$$

ou encore : $x(y) = 1/2 (1 - e^{-2y})$ (HALDANE , 1919)

Schéma illustrant la quasi-liaison entre deux loci situés sur des chromosomes différents . D'après N.T.J.BAILEY .



La quantité y est la distance sur la carte entre deux loci quelconques dont le pourcentage de recombinaison est x . Elle possède les propriétés d'additivité requises.

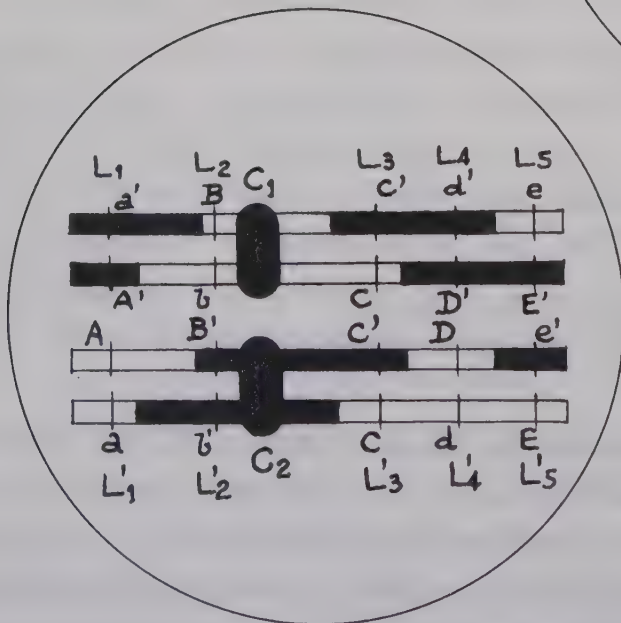
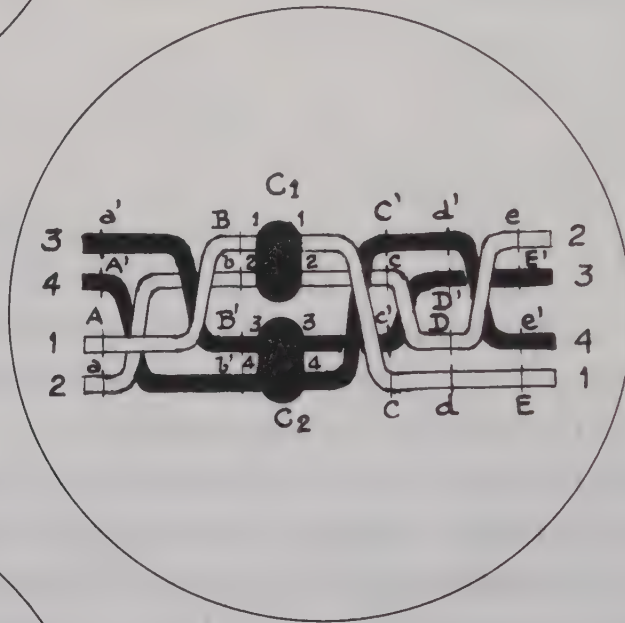
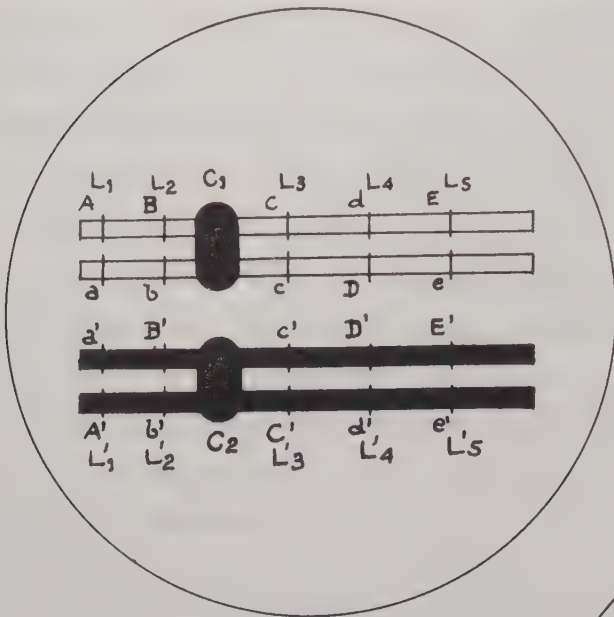
Bien que y ait été introduit par l'intermédiaire d'un artifice purement mathématique, il est possible de l'amener d'une façon plus naturelle par la considération des processus cytologiques impliqués dans le phénomène du crossing-over. " Les cytologistes ont recherché dans les images qu'ils obtenaient de la méiose, si certains aspects ne pourraient pas rendre compte du crossing-over postulé par les généticiens. Ils n'ont vu ni les coupures de chromosomes, ni le recollement des extrémités cassées, mais ils observent couramment, au cours de la prophase hétérotypique, au stade diplotène. c'est-à-dire au stade où les chromosomes sont clivés et où, après le ^{stade} pachytène, ils s'éloignent, des entrecroisements entre chromatides, constituant des boucles. Les noeuds qui limitent ces boucles sont appelés des chiasmas. On en observe pratiquement sur toutes les tétrades de chromatides, un ou plusieurs selon la longueur des chromosomes envisagés. " (N.T.J.BAILEY)

On a pensé, d'abord, que ces chiasmas indiquaient le locus où s'était produit le crossing-over, mais en observant les tétrades de chromatides, à des stades différents, depuis le début du diplotène jusqu'à la métaphase, on s'est aperçu que, plus le stade était avancé, plus les chiasmas étaient proches de l'extrémité du chromosome.

DARLINGTON a fourni de ces observations une interprétation cohérente : la place du chiasma ne représenterait qu'à l'origine, le lieu où se sont produits les deux coupures suivies d'échanges réciproques, mais il n'en serait que la conséquence. Au fur et à mesure du déroulement de la prophase, et à partir du point d'interchange, les deux chromatides remaniés s'éloigneraient l'une de l'autre, si bien que le chiasma se rapprocherait de l'extrémité du chromosome : c'est ce que l'on appelle la terminalisation des chiasmas, la séparation totale des deux chromatides intéressées ne se faisant qu'au début de l'anaphase.

Les résultats expérimentaux nous ont conduits à l'idée que les coupures devaient se produire en deux points strictement homologues ; mais ceci n'est pas contraire aux images que nous venons de schématiser : en effet, les chiasmas apparaissant au ^{stade} diplotène, dès que les paires de chromatides s'éloignent, on peut penser que les cassures et échanges réciproques se

Schéma illustrant le phénomène du crossing-over .



sont produits avant ce stade , c'est-à-dire au plus tard à la fin du stade pachytène , à un moment où les chromatides , soeurs ou non , étaient étroitement serrées les unes contre les autres , parties homologues contre parties homologues .

Ainsi , si cette interprétation est correcte , le crossing-over se produit à la prophase hétérotypique , au stade à quatre chromatides au niveau de deux loci homologues de deux chromatides non soeurs ; s'il en est ainsi , on réalise facilement que lorsqu'un crossing-over se produit , chacun des deux centromères emmène vers les pôles du fuseau de la mitose hétérotypique une chromatide originelle , parentale , et une chromatide remaniée , et non pas deux chromatides entièrement soeurs " (RIZET) .

Le résultat des crossing-over est donc une sorte de " patch-work " ou mosaïque formée de nouvelles combinaisons de gènes , bien que , sauf accident tel qu'inversion ou translocation , l'ordre des loci reste rigoureusement le même .

Pour plus de clarté , dans tout ce qui suivra , nous raisonnerons comme si le crossing-over se produisait entre deux chromosomes et non entre deux chromatides , c'est-à-dire comme s'il s'agissait d'un stade à deux éléments linéaires et non quatre . Ceci n'introduit aucune erreur dans les résultats et simplifie considérablement l'exposé .

L'hypothèse la plus simple , est que les points d'échange entre deux loci quelconques , sont des événements aléatoires , la densité moyenne de ces points restant constante . L'espérance mathématique du nombre de points d'échange (par chromosome) entre deux loci , peut être utilisée comme définition de la distance correspondante sur la carte . Une distance ainsi définie possède évidemment , même en la présence d'interférence , les propriétés additives voulues , puisque l'additivité est une propriété intrinsèque des espérances mathématiques .

Soit x la distance entre deux loci L_1 et L_2 . L'hypothèse que les points d'échange entre L_1 et L_2 surviennent au hasard , revient à dire que le nombre de ces points , apparaissant au cours d'une méiose donnée , suit une distribution de Poisson de paramètre x . La probabilité $p_r(x)$ d'un nombre r de points d'échange entre L_1 et L_2 est alors :

$$p_r(x) = \frac{x^r e^{-x}}{r!} \quad \text{avec } r = 0, 1, 2, \dots$$

Nous ne pouvons , bien évidemment , pas observer ni calculer directement r et x . Nous ne pouvons que calculer une estimation du pourcentage de recombinaisons entre les deux loci . Il faut alors bien insister sur le

fait qu'une recombinaison observable n'apparaît que lorsqu'il survient un nombre impair de points d'échange sur le segment considéré . Il s'en suit immédiatement que , si x_0 est le pourcentage de recombinaisons observées ,

$$x_0 = \sum_{r=0}^{\infty} p_{2r+1}(x) = e^{-x} \sum_{r=0}^{\infty} \frac{x^{2r+1}}{(2r+1)!} = 1/2 (1 - e^{-2x}) \quad (\text{HALDANE})$$

Il existe , en fait , certaines raisons cytologiques de penser que les points d'échange apparaissent selon un processus polarisé , comme nous venons de le voir , débutant au centromère , et s'étendant progressivement de part et d'autre de celui-ci vers les extrémités des bras . Il ne s'agit donc pas , à proprement parler , d'un phénomène purement aléatoire . Cependant , en l'absence d'interférence , il semble logique d'utiliser la distribution de Poisson . Pour de petits segments , il est évident que x_0 et x sont pratiquement équivalents , et aussi , que sur de tels segments , les crossing-over multiples sont négligeables . La fraction de recombinaison et la distance sur la carte , c'est-à-dire l'espérance mathématique du nombre de points d'échange , sont tous deux approximativement égaux à p_1 . La distance sur la carte ainsi définie est mesurée en "morgans" (du nom du généticien T.H.MORGAN) . On utilise souvent aussi le centimorgan , noté cm , et correspondant au pourcentage de la fraction de recombinaison .

De tout ceci il résulte que nous n'avons , actuellement , aucune preuve d'une correspondance physique entre la distance sur la carte , définie comme un concept mathématique , en terme d'espérance mathématique du nombre de points d'échange , et la distance réelle sur le chromosome . De plus , au cours même de la méiose , ces distances varient de façon importante avec la longueur du chromosome au cours des différents stades . Nous n'avons non plus aucune preuve de l'homogénéité du chromosome en ce qui concerne l'aptitude au crossing-over . Il est même au contraire fort probable que le chromosome est hétérogène , composé de régions , ou bandes différentes , plus ou moins riches en points d'échange .

Néanmoins , il est logique de supposer que , schématiquement et approximativement , il existe une assez bonne correspondance entre la distance réelle sur le chromosome et la distance sur la carte . Cette hypothèse a d'ailleurs été confirmée par les nombreuses études effectuées sur *Drosophila Melanogaster* . Dans cette espèce , la longueur totale relative sur la carte , de l'ensemble des chromosomes , correspond approximativement aux longueurs relatives réelles observées cytologiquement .

Finalement , il faut signaler que , bien que nous considérions x et x_0 comme des variables continues , ce sont , très probablement , des variables discrètes . Ceci parce que les loci , quoique de petites dimensions ,

sont finis, et espacés sur des intervalles bien définis . Cependant , comme le nombre des loci est grand , et leurs dimensions petites , un modèle continu semble bien adapté à la plupart des problèmes .

-6/- Théorie élémentaire de l'interférence .

- Introduction .

Au paragraphe précédent , nous avons discuté le problème de l'élaboration d'une carte chromosomique en l'absence d'interférence , c'est-à-dire lorsque l'on suppose que les crossing-over se produisent au hasard et indépendamment les uns des autres . Bien que cette hypothèse soit grossièrement vérifiée dans de nombreux cas , une analyse fine des faits révèle souvent que la formation des crossing-over est un processus non aléatoire . Ceci apparaît d'ailleurs à la fois à l'examen des descendance et à l'étude cytologique de la méiose . Ce phénomène a été baptisé interférence par H.J.MULLER en 1916 , et peut être défini comme l'influence qu'exerce un crossing-over sur les autres crossing-over de la même chromatide . L'interférence habituellement observée est positive, c'est-à-dire qu'un crossing-over tend à empêcher la formation d'autres crossing-over dans son voisinage . Inversement , l'interférence négative , beaucoup plus rare, favoriserait la survenue d'autres crossing-over dans le voisinage de celui qui est étudié .

Il est évident que ce phénomène peut avoir une influence considérable sur les procédés adoptés pour l'élaboration de la carte .

Nous allons d'abord décrire rapidement une des méthodes permettant de déceler l'existence d'un phénomène d'interférence , puis nous essaierons de voir comment estimer son intensité , enfin nous examinerons quelques techniques mathématiques élémentaires conduisant à une carte modifiée , tenant compte de ce phénomène .

- Détection de l'interférence .

Soient trois loci disposés dans l'ordre ABC . Soit x_{10} , la proportion de recombinaisons observées sur le segment AB , x_{20} sur BC , x_{30} sur AC . Nous supposons que x_{10} , x_{20} , x_{30} , ont été déterminés par des observations indépendantes . Soient $\hat{x}_{10}$, $\hat{x}_{20}$, $\hat{x}_{30}$, les estimations obtenues par la méthode du maximum de vraisemblance . Soient , en l'absence d'interférence , $\hat{x}_1$, $\hat{x}_2$, $\hat{x}_3$, les estimations correspondantes de la distance sur la carte . On a :

$$\widehat{x}_i = -1/2 \ln(1 - 2\widehat{x}_{i0}) \quad (i = 1, 2, 3)$$

Si la variance de $\widehat{x}_i$ est v_i , il vient :

$$v_i = \left(\frac{d\widehat{x}_i}{d\widehat{x}_{i0}} \right) \text{var}(\widehat{x}_{i0}) = \frac{\text{var}(\widehat{x}_{i0})}{(1 - 2\widehat{x}_{i0})^2}$$

Les distances réelles sur la carte, x_1, x_2, x_3 , sont additives, de sorte que :

$$x_3 = x_1 + x_2$$

Un bon test de l'existence ou non d'interférence consiste donc à comparer la différence observée $\widehat{x}_3 - (\widehat{x}_1 + \widehat{x}_2)$ à son erreur standard $(v_1 + v_2 + v_3)^{1/2}$

- Estimation de l'intensité de l'interférence . Le concept de "coïncidence"

Nous allons tout d'abord modifier légèrement nos notations, en remplaçant x_{30} par x_{1+20} , et en utilisant le symbole x_{120} pour représenter les recombinaisons survenant simultanément sur AB et BC .

S'il n'y a pas d'interférence, donc si les crossing-over se produisent au hasard, et de façon indépendante, $x_{120} = x_{10} \cdot x_{20}$. Dans le cas d'une interférence positive, la fréquence des doubles crossing-over sera inférieure à la valeur observée dans le cas aléatoire :

$$x_{120} < x_{10} \cdot x_{20}$$

Inversement, dans le cas d'une interférence négative, on observera :

$$x_{120} > x_{10} \cdot x_{20}$$

Par conséquent, le coefficient C dit de "coïncidence" défini par :

$$C = \frac{x_{120}}{x_{10} \cdot x_{20}}$$

est un bon indice de l'intensité de l'interférence . Ce coefficient a été défini par MULLER en 1916 . En l'absence d'interférence, $C = 1$; C est plus petit que 1 lorsque l'interférence est positive, plus grand que 1 lorsqu'elle est négative .

Lorsque C est différent de 1, la formule de TROW, donnée au paragraphe précédent, n'est plus valable ; mais on en obtient facilement une expression modifiée : les recombinaisons sur l'ensemble du segment AC peuvent être envisagées comme l'ensemble des recombinaisons, sur AB seulement, et sur BC seulement . Soient x'_{10} et x'_{20} les fréquences des deux derniers événements . On voit aisément que :

$$x_{10} = x'_{10} + x_{120} \quad \text{et} \quad x_{20} = x'_{20} + x_{120}$$

D'où :

$$x_{1+20} = x'_{10} + x'_{20} = x_{10} + x_{20} - 2x_{120}$$

$$x_{1+20} = x_{10} + x_{20} - 2Cx_{10} \cdot x_{20}$$

qui est la formule analogue à celle de TROW dans le cas de l'interférence .

La théorie mathématique générale de l'interférence est d'une extrême

complexité, et nous ne ferons que passer très rapidement sur cette question, en signalant simplement les idées importantes.

On a pu vérifier, en particulier sur la *Drosophila*, que le coefficient de coïncidence, témoin de l'interférence, n'est pas constant, mais varie, non seulement de chromosome à chromosome, mais aussi, sur un chromosome donné, selon la longueur des segments considérés, et selon leur position par rapport au centromère. Néanmoins, il est possible de trouver des fonctions uniformes, qui donnent des approximations satisfaisantes sur d'assez grandes portions de chromosome. L'équation différentielle de base, reliant la distance sur la carte, x , à la proportion de recombinaisons observées, x_0 , a été donnée par J.B.S. HALDANE, en 1919, sous la forme :

$$x_0 (x + dx) = x_0(x) + dx - 2C_m(x_0) \cdot x_0(x) dx$$

où $C_m(x_0)$ est la "coïncidence marginale" de HALDANE. Il vient alors :

$$\frac{dx_0}{dx} = 1 - 2C_m(x_0) \cdot x_0$$

puisque $\frac{dx}{dx_0} = \frac{1}{1 - 2C_m(x_0) \cdot x_0}$

Nous obtenons la "fonction-carte" :

$$x(x_0) = \int_0^{x_0} \frac{du}{1 - 2C_m(u) \cdot u}$$

où $C_m(u)$ est la coïncidence marginale d'un segment dont la longueur est telle qu'elle conduit à une proportion de recombinaisons égale à u .

Supposons maintenant que les phénomènes d'interférence ne traversent pas le centromère, c'est-à-dire soient indépendants sur les deux bras du chromosome. Chaque bras est considéré comme un continu semi-infini de loci, et peut être représenté par une demi-droite. Tout locus sur le bras, correspond biunivoquement à un point sur la demi-droite, dont la distance à l'origine est t , mesurée relativement à une métrique non encore précisée (l'origine correspond au centromère). Les points d'échange se succèdent sur cette demi-droite. Supposons que la distance u , entre deux points d'échange consécutifs, ait une fonction de fréquence $f(u)$, et une fonction de répartition :

$$F(u) = \int_0^u f(u) du$$

Soient $f_r(t)$ et $F_r(t)$, les fonctions de fréquence et de répartition de la distance du r^{e} point d'échange à partir de l'origine. $F_r(t)$ est donc la probabilité que le r^{e} point d'échange soit à la distance $t' < t$ de l'origine. D'où, si $p_r(t)$ est la probabilité de r points d'échange sur

le segment $[0, t]$:

$$p_n(t) = F_n(t) - F_{n+1}(t)$$

Posons :

$$p(\lambda, t) = \sum_{n=0}^{\infty} \lambda^n p_n(t)$$

$x(t)$, distance précédemment considérée, est le nombre moyen d'échanges sur $[0, t]$, soit :

$$x(t) = \sum_{n=0}^{\infty} n \cdot p_n(t) = \frac{\partial p}{\partial \lambda} \quad (\lambda = 1)$$

$x_0(t)$ est la proportion de recombinaisons c'est-à-dire la probabilité d'un nombre impair d'échanges :

$$x_0(t) = \sum_{n=0}^{\infty} p_{2n+1}(t) = 1/2 [1 - p(-1, t)]$$

On utilise alors la transformée de LAPLACE et ses propriétés relatives aux convolutions. On obtient les résultats suivants :

soit $f^*(s)$ la transformée de $f(u)$

soit $x^*(s)$ la transformée de $x(t)$

soit $x_0^*(s)$ la transformée de $x_0(t)$

alors :

$$x^*(s) = \frac{f^*(s)}{s[1-f^*(s)]}$$

$$x_0^*(s) = \frac{f^*(s)}{s[1+f^*(s)]}$$

Quand on connaît $f(u)$, on peut donc connaître $x(t)$ et $x_0(t)$. S'il n'y a pas d'interférence, on admet que les points d'échange suivent une loi de POISSON :

$$f(u) = ke^{-ku}$$

On retrouve alors la formule de HALDANE .

- Exemple de métrique .

Nous citerons seulement rapidement la métrique de OWEN :

$$f(u) = 4ue^{-2u}$$

c'est la métrique $1/4 \chi^2_4$

tous les calculs peuvent être généralisés à une métrique $1/2m \chi^2_{2m}$

Il vient alors :

$$p_n(t) = \left\{ \frac{(2t)^{2n}}{(2n)!} + \frac{(2t)^{2n+1}}{(2n+1)!} \right\} e^{-2t}$$

$$x(t) = -1/4 + t + 1/4 \cdot e^{-4t}$$

$$x_0(t) = 1/2 - 1/2 \cdot e^{-2t} (\cos 2t + \sin 2t)$$

ceci, en supposant les chromosomes infinis. Si l'on veut tenir compte de la longueur finie des bras chromosomiques, il faut calculer la probabilité de n'importe quelle configuration sur $[0, t]$, conditionnée par le fait qu'il n'y ait pas de points d'échange sur $[t, r]$ avec $r \rightarrow \infty$.

-7/- Combien faudrait-il connaître de marqueurs pour pouvoir immédiatement situer un locus quelconque ?

Il est intéressant de se demander combien de marqueurs seraient nécessaires pour couvrir l'ensemble du génome de telle façon qu'une liaison concernant un locus quelconque puisse à coup sûr être directement mise en évidence . J.MOHR , très récemment , a posé ce problème sous la forme des deux questions suivantes :

- a)- Quelle est la distribution la plus probable de chromosomes possédant 1,2,3 marqueurs ou davantage ?
- b)- Quelle est la probabilité qu'un locus quelconque soit situé suffisamment près (moins de 20% d'unités de recombinaison) d'un système de marqueurs connu pour que sa liaison avec ce système soit directement décelable ? (la limite 20% est choisie arbitrairement) .

La deuxième question peut aussi se poser sous la forme : quelle est la probabilité pour qu'un locus quelconque soit situé à moins de 20 unités de recombinaison de l'un quelconque des systèmes marqueurs connus ?

- Méthode .

Elle est due à J.MOHR . On a considéré différentes possibilités de longueur totale du génome : 1000 , 1667 , 2857 , 5000 centimorgans . La valeur la plus probable actuellement , confirmée par les observations cytologiques de FORD et HAMILTON , (1956) , et FORD (1962) , est d'environ 2640 unités . Les longueurs relatives des chromosomes 1 à 22 seraient : 95 , 87 , 74 , 68 , 62 , 58 , 50 , 47 , 47 , 46 , 46 , 45 , 37 , 34 , 34 , 31 , 31 , 26 , 23 , 22 , 19 et 18 pour 1000 . Ces résultats sont basés sur la mesure des chromosomes effectuée par TJIO et PUCK (1958) .

Grâce à une calculatrice électronique, travaillant sur une série de 90000 nombres de trois chiffres, tirés au hasard , J.MOHR a pu effectuer une étude statistique permettant , par extrapolation , de répondre approximativement aux questions posées . Celles-ci ont été examinées dans les cas de 14 , 17 , 30 , 45 , 60 , 75 et 90 systèmes marqueurs .

Pour chaque nombre donné de marqueurs , on a cherché à déterminer la surface du génome ainsi délimitée . Par exemple , dans le cas de 14 marqueurs , la machine les distribue au hasard entre les différents chromosomes dont les longueurs respectives sont représentées au moyen de nombres tirés au hasard . Un millier de tirages de ce type ont été effectués . A chaque tirage la machine compte le nombre de chromosomes possédant 0 , 1 , 2 , 3 marqueurs etc ... et calcule la moyenne de ces distributions

pour les mille tirages .

On examine ensuite , pour chaque tirage , de quelle façon les 14 marqueurs couvrent le génome . Un système marqueur donné, couvre au maximum 40 unités , (20 unités de part et d'autre) . La longueur couverte est , en moyenne , inférieure à 40 unités, en raison :

- 1/- de la brièveté de certains chromosomes
- 2/- du chevauchement possible des zones correspondant à deux marqueurs voisins .
- 3/- de la position de certains marqueurs à moins de 20 unités des extrémités des chromosomes .

La machine mesure , pour chaque locus marqueur , soit 20 unités de part et d'autre , soit la distance à l'extrémité du chromosome, si celle-ci est inférieure à 20 . On calcule enfin la moyenne de la zone marquée totale pour les 1000 tirages .

- Résultats .

Le tableau 1 donne la distribution des chromosomes non marqués , marqués une , deux , trois fois et plus , pour des nombres divers de marqueurs allant de 14 à 90 .

Il apparaît que , avec 14 marqueurs , à peu près la moitié des autosomes restent non marqués , alors que 7 environ ne possèdent qu'un marqueur , et trois , plus d'un marqueur .

Avec 30 marqueurs , ce qui représente à peu près le double du nombre de marqueurs dont nous disposons actuellement chez l'homme (tableau 3) , il resterait encore plus de 6 chromosomes non marqués en moyenne , et 7 ne seraient marqués qu'une seule fois .

Avec 90 marqueurs , il ne resterait qu'un chromosome non marqué , 2 à 3 ne possèderaient qu'un marqueur , et le reste des chromosomes serait marqué par plus d'un système .

Le tableau 2 donne la probabilité d'une liaison décelable directement , c'est-à-dire la probabilité pour qu'un locus donné soit situé à moins de 20 unités d'un marqueur quelconque .

Pour une longueur chromosomique totale de 2857 (voisine de la situation réelle) , et pour 14 systèmes marqueurs , cette probabilité est de 0,17 ; elle passe à 0,32 pour 30 marqueurs et 0,68 pour 90 marqueurs .

Il est donc éminemment nécessaire et urgent de trouver de nombreux nouveaux marqueurs, pour augmenter nos chances de déceler des liaisons . Un gain , même modeste , augmenterait considérablement ces chances . Il faut également noter que , même avec 90 marqueurs , ce qui apparaît actuellement comme utopique, le génome ne serait qu'impartiellement marqué.

Table I

Average distribution of unmarked, doublemarked etc. chromosomes for various numbers of marker systems

Average No. of unmarked (0), singlemarked (1), double- marked (2) etc. chromosomes	No. of marker systems						
	14	17	30	45	60	75	90
0	11.91	10.53	6.57	3.97	2.49	1.64	1.09
1	7.06	7.35	7.05	5.73	4.33	3.32	2.51
2	2.31	3.00	4.56	4.96	4.80	4.11	3.47
3	0.60	0.88	2.29	3.43	3.83	3.73	3.50
4	0.10	0.20	0.99	1.97	2.64	3.07	3.11
5	0.02	0.03	0.35	1.07	1.72	2.21	2.51
6	0.01		0.13	0.51	1.04	1.51	1.88
7			0.05	0.22	0.56	1.02	1.39
8			0.01	0.08	0.30	0.61	0.98
9				0.05	0.16	0.38	0.60
10				0.01	0.07	0.20	0.38
11				0.01	0.03	0.10	0.27
12					0.02	0.06	0.15
13					0.01	0.03	0.07
14						0.01	0.05
15						0.01	0.02
16							0.01
17							0.01

Table II

Probability of a chosen locus being within 20 units distance from one or other of the known systems, for various postulated total lengths of the genome, and for various numbers of marker systems

Postulated total length of genome	No. of marker systems						
	14	17	30	45	60	75	90
1000	0.35	0.41	0.60	0.73	0.82	0.87	0.91
1667	0.26	0.30	0.46	0.60	0.71	0.78	0.84
2857	0.17	0.20	0.32	0.44	0.54	0.62	0.68
5000	0.10	0.12	0.21	0.29	0.37	0.44	0.50

CHAPITRE II

- LES METHODES DE CALCUL DU POURCENTAGE DE RECOMBINAISONS .

Au début de ce chapitre , nous rappellerons brièvement , avec un exemple , le principe de la méthode du maximum de vraisemblance , méthode statistique à laquelle nous ferons allusion à plusieurs reprises au cours des pages qui suivront .

- METHODE DU MAXIMUM DE VRAISEMBLANCE .

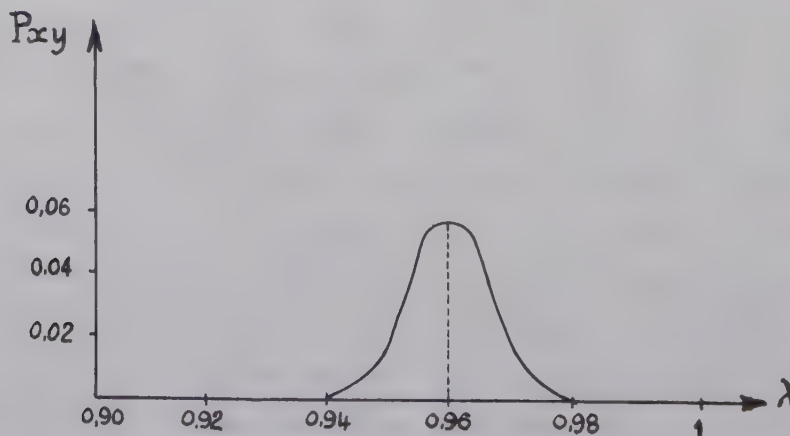
-1/- Il s'agit de trouver quelle est la probabilité d'obtenir un échantillon tel que celui observé . La probabilité p_{xy} d'obtenir x évènements de type A et y évènements de type B , sur un total de $n = x + y$ évènements , est , si p est la probabilité de l'évènement A , et q celle de l'évènement B :

$$(1) \quad p_{xy} = p^x q^y \frac{n!}{x!y!}$$

Si l'on prend comme exemple le groupe sanguin Lutheran , l'évènement A est la mise en évidence d'un sujet $Lu_a Lu_a$ ou $Lu_a Lu_b$, l'évènement B , celle d'un sujet $Lu_b Lu_b$. Soit λ la fréquence du gène Lu_b , $1 - \lambda$ celle du gène Lu_a . λ^2 , fréquence des sujets $Lu_b Lu_b$, est égal à 536 , et $1 - \lambda^2$, fréquence des sujets $Lu_a Lu_a + Lu_a Lu_b$, est égal à 46 , dans un échantillon tiré au hasard , de 582 sujets . Mais λ , fréquence observée , et non probabilité , n'est pas connu avec certitude . D'après la formule (1) on a :

$$p_{xy} = (1 - \lambda^2)^{46} (\lambda^2)^{536} \frac{582!}{46! 536!}$$

Différentes valeurs hypothétiques de λ peuvent être essayées et les valeurs correspondantes de p_{xy} calculées . Nous obtenons ainsi un graph de p_{xy} en fonction de λ , pour des valeurs de λ comprises entre 0,9 et 1 puisque le λ observé ici était égal à 0,96 :



Le maximum de p_{xy} apparaît exactement pour notre estimation de $\lambda = 0,96$.

En passant aux logarithmes, ce qui simplifie les calculs, il vient :

$$\log p_{xy} = L = \ln\left(\frac{582!}{46! 536!}\right) + 46 \ln(1 - \lambda^2) + 536 \ln \lambda^2$$

et en dérivant par rapport à λ :

$$D_{\lambda} L = \frac{46 (-2\lambda)}{1 - \lambda^2} + \frac{536 \cdot 2\lambda}{\lambda^2} = U$$

Si L est maximum, $D_{\lambda} L = U = 0$. C'est-à-dire, si l est la valeur de λ pour laquelle L est maximum :

$$\frac{46 (-2l)}{1 - l^2} + \frac{536 \cdot 2l}{l^2} = 0$$

puisque l est différent de 0 ; d'où :

$$- 46 l^2 + 536 (1 - l^2) = 0$$

soit : $l^2 = 536 / 582$ c'est-à-dire $l = 0,96$

-2/- Pour trouver l'erreur de la meilleure estimation l de λ , nous procéderons de la façon suivante :

Calculons $I = - D_{\lambda}^2 L = - D_{\lambda} U$, dérivée seconde de L par rapport à λ .
Après simplification, il vient :

$$I = \frac{92}{1 - \lambda^2} + \frac{184}{(1 - \lambda^2)^2} + \frac{1072}{\lambda^2}$$

Remplaçons λ par son estimation $l = 0,96$, ceci nous donne :

$$I = 2,99 \times 10^{-4}$$

La variance v de l est approximativement égale à :

$$v = 1/I$$

et son écart-type s :

$$s = 1/\sqrt{I} = 0,0058$$

Ainsi, la vraie valeur de λ est comprise entre $l - 2s = 0,948$ et $l + 2s = 0,972$ avec 95% de sécurité. La quantité I a été appelée par R.A.FISHER "la quantité d'information sur λ , contenue dans l'échantillon étudié".
 l est l'estimation la plus vraisemblable. Elle peut être définie comme la valeur de λ pour laquelle la probabilité d'obtenir l'échantillon observé est maxima. (Notons bien que nous ne disons pas que l est la valeur la plus probable du paramètre inconnu λ).

Nous n'exposerons pas ici la méthode des scores- u de R.A.FISHER, qui n'est plus utilisée actuellement sous sa forme initiale. nous donnerons seulement un bref aperçu de ses inconvénients dans la conclusion de ce chapitre. Nous nous bornerons à l'étude des trois méthodes les

plus usuelles : la méthode des paires de germains de L.S.PENROSE , la méthode du rapport de probabilités de C.A.B.SMITH , et la méthode séquentielle de N.E.MORTON .

Pour simplifier les notations , nous utiliserons à partir de maintenant le symbole x , et non plus x_0 , pour désigner le pourcentage de recombinaisons observées .

- METHODE DES PAIRES DE GERMAINS DE L.S.PENROSE .

Cette méthode ne nécessite pas la connaissance des génotypes des parents . Elle est basée sur l'étude des paires indépendantes en probabilité que l'on peut former à partir des enfants d'une même fratrie . Le principe est le suivant :

-1/- Paires dans le cas où l'on considère deux allèles , A et a , A étant dominant sur a .

Les différentes paires de frères et soeurs que l'on peut former appartiennent à l'un des trois types possibles suivants :

- (1) $\left\{ \begin{array}{l} \text{- Dominant-Dominant (DD)} \\ \text{- Dominant-Récessif (DR)} \\ \text{- Récessif-Récessif (RR)} \end{array} \right.$

Considérons par exemple un mariage :

(2) $Aa \times Aa$

dont la fréquence dans la population est :

$$2pq \times 2pq = 4p^2q^2$$

Ce mariage peut produire des enfants de trois types :

$$AA - Aa - aa$$

dans des proportions respectives :

$$1/4 - 1/2 - 1/4$$

soit $3/4$ de D et $1/4$ de R .

Il en résulte que , dans la population entière , les fréquences des paires (1) seront données par le développement du carré symbolique :

$$(3) \quad 4p^2q^2(3/4D + 1/4R)^2 = 9/4p^2q^2(DD) + 3/2p^2q^2(DR) + 1/4p^2q^2(RR)$$

De calculs analogues on déduit le tableau de la page suivante d'où l'on tire :

$$(4) \quad T_{DD} + T_{DR} + T_{RR} = 1$$

$$(5) \quad T_{DD} = 1 - q^2 - 1/4pq^2(3 + q) = 1 - q^2 - 1/2T_{DR}$$

$$(6) \quad T_{DR} = 1/2pq^2(3 + q)$$

$$(7) \quad T_{RR} = 1/4q^2(1 + q)^2 = q^2 - 1/2T_{DR}$$

$$(8) \quad T_{DD} + T_{RR} - T_{DR} = 1 - 2T_{DR} = 1 - pq^2(3 + q)$$

Tableau I

types de mariages et fréquences correspondantes		fréquences des paires de frères		
		D-D	D-R	R-R
AA x AA	p^4	p^4	0	0
AA x Aa	$4p^3q$	$4p^3q$	0	0
AA x aa	$2p^2q^2$	$2p^2q^2$	0	0
Aa x Aa	$4p^2q^2$	$9/4p^2q^2$	$3/2p^2q^2$	$1/4p^2q^2$
Aa x aa	$4pq^3$	pq^3	$2pq^3$	pq^3
aa x aa	q^4	0	0	q^4
Totaux	1	T_{DD}	T_{DR}	T_{RR}

-2/- Paires dans le cas où l'on considère deux paires d'allèles : (A,a) (B,b), A étant dominant sur a et B sur b.

Désignons par C_P (caractère principal), et C_M (caractère marqueur), les caractères conditionnés par ces paires de gènes. Nous pouvons grouper les seize paires de germains possibles comme il est indiqué dans le tableau ci-dessous :

Tableau II

		Paires dans lesquelles C_P	
		a même expression	a des expressions différentes
paires dans lesquelles C_M	a même expression	1 $D_P D_M - D_P D_M$ $D_P R_M - D_P R_M$ $R_P D_M - R_P D_M$ $R_P R_M - R_P R_M$	2 $D_P D_M - R_P D_M$ $D_P R_M - R_P R_M$ $R_P D_M - D_P D_M$ $R_P R_M - D_P R_M$
	a des expressions différentes	3 $D_P D_M - D_P R_M$ $D_P R_M - D_P D_M$ $R_P D_M - R_P R_M$ $R_P R_M - R_P D_M$	4 $D_P D_M - R_P R_M$ $D_P R_M - R_P D_M$ $R_P D_M - D_P R_M$ $R_P R_M - D_P D_M$

La case n° 1 contient les paires dans lesquelles les deux caractères sont semblables : paires CC.

La case n°4 contient les paires dans lesquelles les deux caractères diffèrent : paires $\Delta\Delta$.

La case n°3 contient les paires dans lesquelles les deux germains sont semblables pour C_2 et différents pour C_1 : paires $C\Delta$.

Enfin, la case n°2 contient les paires ΔC , différentes pour C_2 et semblables pour C_1 .

Nous poserons, dans la population générale :

$$\begin{aligned} P_n(CC) &= \omega_1 & P_n(\Delta C) &= \omega_2 \\ P_n(C\Delta) &= \omega_3 & P_n(\Delta\Delta) &= \omega_4 \end{aligned}$$

a)- cas de l'indépendance : L_1 et L_2 sont sur deux chromosomes différents.

Si les paires d'allèles (A,a) au locus L_1 et (B,b) au locus L_2 sont indépendantes, et si les fréquences de ces gènes sont, respectivement

$$p_1, q_1, p_2, q_2$$

nous avons, en utilisant les résultats du tableau I, et en remarquant que, par exemple, $P_n(D_1 R_1) = P_n(R_1 D_1) = 1/2 T_{D_1 R_1}$

$$\omega_1 = T_{D_1 D_1} \times T_{D_2 D_2} + T_{D_1 D_1} \times T_{R_1 R_1} + T_{R_1 R_1} \times T_{D_2 D_2} + T_{R_1 R_1} \times T_{R_2 R_2}$$

$$\omega_1 = (T_{D_1 D_1} + T_{R_1 R_1})(T_{D_2 D_2} + T_{R_2 R_2})$$

$$(9) \quad \omega_1 = (1 - T_{D_1 R_1})(1 - T_{D_2 R_2})$$

$$\text{De même,} \quad \omega_2 = T_{D_1 R_1}(1 - T_{D_2 R_2})$$

$$(10) \quad (11) \quad (12) \quad \omega_3 = T_{D_1 R_1}(1 - T_{D_2 R_2})$$

$$\omega_4 = T_{D_1 R_1} \times T_{D_2 R_2}$$

b)- cas où L_1 et L_2 sont liés.

Les fréquences dans la population générale, supposée en équilibre, des neuf génotypes possibles, sont les mêmes, que les loci soient liés ou non. Ces fréquences sont données par la matrice :

$$\begin{bmatrix} p_1^2 p_2^2 & 2p_1^2 p_2 q_2 & p_1^2 q_2^2 \\ 2p_1 q_1 p_2^2 & 4p_1 q_1 p_2 q_2 & 2p_1 q_1 q_2^2 \\ q_1^2 p_2^2 & 2q_1^2 p_2 q_2 & q_1^2 q_2^2 \end{bmatrix}$$

D'autre part, on vérifie très facilement que parmi tous les mariages possibles, seuls les quatre suivants auront une descendance dépendant de la force de la liaison :

$$(1) \quad \begin{pmatrix} a \\ b \end{pmatrix} \parallel \begin{pmatrix} a \\ b \end{pmatrix} \times \begin{pmatrix} A \\ B \end{pmatrix} \parallel \begin{pmatrix} a \\ b \end{pmatrix}$$

$$(2) \quad \begin{pmatrix} a \\ b \end{pmatrix} \parallel A \times \begin{pmatrix} a \\ b \end{pmatrix} \parallel B$$

$$(3) \quad \begin{pmatrix} a \\ b \end{pmatrix} \parallel a \times \begin{pmatrix} a \\ b \end{pmatrix} \parallel B$$

$$(4) \quad \begin{pmatrix} a \\ b \end{pmatrix} \parallel A \times \begin{pmatrix} a \\ b \end{pmatrix} \parallel B$$

Etudions, par exemple, les mariages du type (1), en supposant qu'il y ait phase. Si x désigne le pourcentage de crossing-over, on a :

$$\begin{aligned} & \left(\frac{a}{b} \right) \left[\left(\frac{1-x}{2} \right) \left(\frac{A}{B} \right) + \left(\frac{1-x}{2} \right) \left(\frac{a}{b} \right) + \frac{x}{2} \left(\frac{A}{b} \right) + \frac{x}{2} \left(\frac{a}{B} \right) \right] = \\ & = \left(\frac{1-x}{2} \right) \left(\frac{a}{b} \parallel \frac{A}{B} \right) + \left(\frac{1-x}{2} \right) \left(\frac{a}{b} \parallel \frac{a}{b} \right) + \frac{x}{2} \left(\frac{a}{b} \parallel \frac{A}{b} \right) + \frac{x}{2} \left(\frac{a}{b} \parallel \frac{a}{B} \right) \end{aligned}$$

En effectuant le carré symbolique de cette dernière expression, nous obtiendrons les différents types de paires de germains possibles et leurs fréquences. Après réduction des termes semblables, il reste :

$$(13) \quad \left[(1-x)^2 + \frac{x^2}{2} \right] (CC) + x(1-x)(CA) + x(1-x)(\Delta C) + \frac{x^2}{2} (\Delta \Delta)$$

L'expression, dans le cas de la répulsion, s'obtient en changeant x en $1-x$ et inversement. Il vient alors :

$$(14) \quad \left[x^2 + \frac{(1-x)^2}{2} \right] (CC) + x(1-x)(CA) + x(1-x)(\Delta C) + \frac{(1-x)^2}{2} \Delta \Delta$$

La probabilité d'obtenir, dans la population générale, un mariage du type (1) est, aussi bien dans le cas de phase que dans celui de répulsion, $4p_1 q_1^3 p_2 q_2^3$; il en résulte que l'on a :

$$P_2 [(1) \text{ et } (CC)] = 6 p_1 q_1^3 p_2 q_2^3 [x^2 + (1-x)^2]$$

$$P_2 [(1) \text{ et } (CA)] = 8 p_1 q_1^3 p_2 q_2^3 x(1-x)$$

$$P_2 [(1) \text{ et } (\Delta C)] = 8 p_1 q_1^3 p_2 q_2^3 x(1-x)$$

$$P_2 [(1) \text{ et } (\Delta \Delta)] = 2 p_1 q_1^3 p_2 q_2^3 [x^2 + (1-x)^2]$$

Soit N le nombre des paires de germains, indépendantes, extraites au hasard de la population générale. Parmi ces paires, certaines pourront être nées de mariages du type (1). Nous désignerons par :

$$X_{11}, X_{12}, X_{13}, X_{14}$$

les variables aléatoires représentant respectivement les nombres possibles de paires CC , CA , ΔC , $\Delta \Delta$. Nous avons :

$$E_1 = E [(X_{11} + X_{14}) - (X_{12} + X_{13})]$$

$$= 8 p_1 q_1^3 p_2 q_2^3 [x^2 + (1-x)^2 - 2x(1-x)] N = 16 p_1 q_1^3 p_2 q_2^3 \left[2 \left(x - \frac{1}{2} \right)^2 \right] N$$

En étudiant les trois autres types de mariages on vérifiera que :

$$E_2 = 16 p_1^2 q_1^2 p_2 q_2 (x - \frac{1}{2})^2 N$$

$$E_3 = 16 p_1 q_1^3 p_2^2 q_2^2 (x - \frac{1}{2})^2 N$$

$$E_4 = 16 p_1^2 q_1^2 p_2^2 q_2^2 (x - \frac{1}{2})^2 [(x - \frac{1}{2})^2 + \frac{1}{2}] N$$

Posons alors :

$$X_{c1} = X_{11} + X_{21} + X_{31} + X_{41}$$

$$X_{c2} = X_{12} + X_{22} + X_{32} + X_{42} \quad \text{etc...}$$

et

$$Lx = E_1 + E_2 + E_3 + E_4 = E [(X_{c1} + X_{c4}) - (X_{c2} + X_{c3})]$$

nous aurons finalement :

$$(15) \quad Lx = 16 p_1 q_1^2 p_2 q_2^2 [\frac{1}{2} (1+q_1)(1+q_2) + p_1 p_2 (x - \frac{1}{2})^2] (x - \frac{1}{2})^2 N$$

S'il n'y a pas liaison, $X = 1/2$ et $Lx = 0$.

Considérons maintenant l'échantillon complet des N paires de germains.

Soient X_1 , X_2 , X_3 et X_4 , les nombres possibles de paires CC, CΔ, ΔC et ΔΔ respectivement. Posons :

$$(16) \quad L = E [(X_1 + X_4) - (X_2 + X_3)]$$

Nous pouvons distinguer, dans l'ensemble L , deux parties complémentaires :

- L_x , correspondant aux mariages (1), (2), (3), (4).
- L_o , correspondant aux autres types de mariages.

Nous avons évidemment :

$$(17) \quad L = L_o + L_x$$

et comme $L_x = 0$ pour $x = 1/2$, nous pouvons dire aussi que L_o est la valeur de L en l'absence de liaison. Il s'en suit :

$$L_o = N [(\omega_1 + \omega_4) - (\omega_2 + \omega_3)]$$

La valeur des ω_i étant donnée par les relations (9), (10), (11) et (12). On trouve facilement :

$$(18) \quad L_o = N [1 - p_1 q_1^2 (3 + q_1)] [1 - p_2 q_2^2 (3 + q_2)]$$

- Application pratique de la méthode de L.S.PENROSE.

De l'ensemble des données familiales, nous tirons une estimation de L . Connaissant p_1 , q_1 , p_2 et q_2 , nous pouvons calculer L_o . Par différence, nous obtiendrons une estimation de L_x qui nous permettra d'évaluer x .

- Extension de la méthode des paires de germains.

Dans une série de mémoires, L.S.PENROSE a étendu, ou perfectionné, sa méthode et montré sa similitude avec celle des scores "U" de R. FISHER. FINNEY a étudié la perte d'information qui résulte des conditions de son emploi et conclut qu'elle ne peut être utile que pour

avoir une indication sur l'existence d'une liaison.

- Exemple de détection d'une liaison entre deux loci.

. Méthode de 2x2

Pour mettre en évidence la liaison, et non mesurer son intensité, on peut utiliser la méthode suivante :

Reprenons le tableau II; dans le cas de l'indépendance, les espérances mathématiques de chacune des cases sont respectivement :

$$a_1 = Nw_1, \quad a_2 = Nw_2, \quad a_3 = Nw_3, \quad a_4 = Nw_4$$

Nous vérifions immédiatement, d'après (9), (10), (11) et (12) que :

$$a_1 = \frac{a_1 + a_3}{N} (a_1 + a_2) \quad a_3 = \frac{a_1 + a_3}{N} \quad \text{etc...}$$

Il en résulte que, si nous considérons le tableau III ci-dessous,

		Paires dans lesquelles C P		TOTAUX
		a même expression	a des expressions $\neq$	
Paires dans lesquelles C M	a même expression	CC a_1	ΔC a_2	$a_1 + a_2$
	a des expressions différentes	C Δ a_3	$\Delta\Delta$ a_4	$a_3 + a_4$
	TOTAUX	$a_1 + a_3$	$a_2 + a_4$	N

comme une table 2x2 où les fréquences marginales sont fixées, la valeur du χ^2 est nulle.

Ceci posé, si N est de l'ordre de 100, et si n_1, n_2, n_3, n_4 , sont les fréquences observées, la valeur de :

$$\chi^2 = N \frac{(n_1 n_4 - n_2 n_3)^2}{(n_1 + n_2)(n_1 + n_3)(n_2 + n_4)(n_3 + n_4)}$$

donnera une indication sur la probabilité d'une liaison.

- Remarque fondamentale.

Rappelons cependant ce qui a été dit au début de ce paragraphe, à savoir qu'il n'est pas correct de déduire $\frac{n(n-1)}{2}$ paires de germains d'une fratrie de n enfants. Nos démonstrations supposent que l'échantillon est formé de N paires indépendantes. Lorsque les génotypes des parents sont connus, les paires de frères et soeurs sont sans

corrélation, mais il n'en est plus ainsi lorsque les génotypes parentaux sont partiellement inconnus. (C.A.B. SMITH).

• Méthode de 3x3

Nous pouvons envisager une autre répartition des paires de germains, comme l'indique le tableau IV ci-dessous :

		Caractère C _M			
		D _M D _M	D _M R _M	R _M R _M	Totaux
Caractère C _P	D _P D _P	a ₁₁	a ₁₂	a ₁₃	B ₁
	D _P R _P	a ₂₁	a ₂₂	a ₂₃	B ₂
	R _P R _P	a ₃₁	a ₃₂	a ₃₃	B ₃
Totaux		A ₁	A ₂	A ₃	N

les a_{ij} représentent les espérances mathématiques de chacune des combinaisons phénotypiques, lorsque le nombre des paires de l'échantillon observé est N. Posons :

$$Pr(D_1D_1) = \pi_1, \quad Pr(D_1R_1) = \pi_2, \quad Pr(R_1R_1) = \pi_3$$

$$Pr(D_2D_2) = \nu_1, \quad Pr(D_2R_2) = \nu_2, \quad Pr(R_2R_2) = \nu_3$$

avec :

$$\pi_1 + \pi_2 + \pi_3 = 1, \quad \nu_1 + \nu_2 + \nu_3 = 1$$

Si les deux loci sont indépendants on a, par exemple,

$$Pr(D_1D_1; D_2R_2) = Pr(D_1D_1) \cdot Pr(D_2R_2) = \pi_1 \cdot \nu_2$$

d'où : $a_{12} = N\pi_1\nu_2$

On vérifie de même les égalités suivantes :

$$a_{11} = N\pi_1\nu_1$$

$$a_{12} = N\pi_1\nu_2$$

$$a_{13} = N\pi_1\nu_3$$

$$a_{21} = N\pi_2\nu_1$$

$$a_{22} = N\pi_2\nu_2$$

$$a_{23} = N\pi_2\nu_3$$

$$a_{31} = N\pi_3\nu_1$$

$$a_{32} = N\pi_3\nu_2$$

$$a_{33} = N\pi_3\nu_3$$

$$A_1 = N\nu_1$$

$$A_2 = N\nu_2$$

$$A_3 = N\nu_3$$

$$B_1 = N\pi_1$$

$$B_2 = N\pi_2$$

$$B_3 = N\pi_3$$

qui montrent que $a_{ij} = B_i \frac{A_j}{N}$

Par conséquent, en l'absence de liaison, le χ^2 calculé sur les a_{ij} est nul.

Sa valeur, calculée sur les n_{ij} observés, donnera une indication sur l'existence d'une hypothétique liaison.

- METHODE DU RAPPORT DE PROBABILITES DE H.B.S. HALDANE ET C.A.B. SMITH.

Soient L_1 et L_2 les deux loci dont nous voulons étudier la liaison éventuelle. Nous appellerons x le pourcentage de recombinaisons entre L_1 et L_2 , et F , l'échantillon des n familles $F_1, F_2, \dots, F_n$ susceptibles d'apporter une information. Pour calculer $P(F_1; x)$, probabilité d'obtenir la famille F_1 dans l'hypothèse H_1 où L_1 et L_2 sont liés, il est nécessaire d'envisager toutes les combinaisons génétiques possibles, s'accordant avec les phénotypes observés. Chacune de ces k combinaisons a une probabilité $p_1, \dots, p_k$. $P(F_1; x)$ est la somme de ces probabilités élémentaires de distributions mutuellement exclusives

$$P(F_1; x) = \sum_{i=1}^k p_i$$

Cette probabilité s'exprime sous la forme d'un polynôme de degré variable, et parfois élevé, en x . De la même façon, $P(F_1; 1/2)$, est la probabilité d'obtenir la famille F_1 dans l'hypothèse H_0 (hypothèse nulle) où les loci L_1 et L_2 sont indépendants. Nous appellerons $P(F; x)$, la probabilité d'obtenir l'ensemble F des n familles dans l'hypothèse H_1 , c'est-à-dire le produit des n probabilités $P(F_i; x)$.

$$P(F; x) = \prod_{i=1}^n P(F_i; x)$$

De même :

$$P(F; 1/2) = \prod_{i=1}^n P(F_i; 1/2)$$

est la probabilité d'obtenir l'échantillon F dans l'hypothèse H_0 d'indépendance. Nous définirons pour chaque famille la quantité $\lambda(F_i; x)$, comme le rapport de la probabilité $P(F_i; x)$ pour un pourcentage x de recombinaisons, à la probabilité $P(F_i; 1/2)$ dans le cas d'indépendance.

$$\lambda(F_i; x) = \frac{P(F_i; x)}{P(F_i; 1/2)}$$

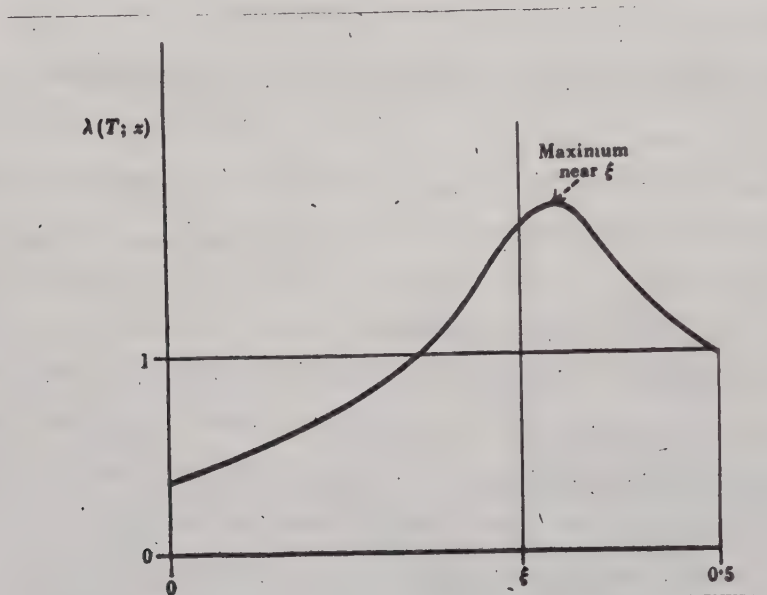
Par généralisation à l'échantillon tout entier, nous définissons $\lambda(F; x)$ comme le produit des $n \lambda(F_i; x)$:

$$\lambda(F; x) = \prod_{i=1}^n \lambda(F_i; x)$$

Lorsque l'on fait varier x de 0 à $1/2$, $\lambda(F; x)$, proportionnel à

$P(F;x)$, passe par un maximum pour une valeur de x proche de ξ , vraie valeur du pourcentage de recombinaisons.

D'autre part, $\lambda(F;1/2) = 1$ par définition. La courbe représentative des variations de $\lambda(F;x)$ en fonction de x a donc l'aspect schématisé ci-dessous :



Un maximum très net de cette courbe pour une valeur ξ' de x , permet de penser qu'il y a liaison et de donner une estimation, ξ' , du pourcentage réel de recombinaisons, ξ .

Nous définirons encore pour chaque famille la quantité $\Lambda(F_i)$, comme la valeur moyenne de $\lambda(F_i; x)$ dans l'intervalle $0 \leq x \leq 1/2$. Soit:

$$\Lambda(F_i) = 2 \int_0^{1/2} \lambda(F_i; x) dx$$

De la même façon :

$$\Lambda(F) = 2 \int_0^{1/2} \lambda(F; x) dx$$

Nous avons maintenant le choix entre plusieurs méthodes d'analyse :

a) - Méthode des probabilités a posteriori ou méthode inverse.

Nous pouvons considérer les deux loci L_1 et L_2 comme un échantillon tiré au hasard parmi tous les couples de loci possibles. x constitue alors une variable aléatoire possédant une certaine distribution "a priori". Entre 0 et $1/2$ cette distribution est continue; nous la désignerons par $\phi(x)dx$. La valeur $x = 1/2$ a une probabilité définie

par Φ de sorte que l'on a :

$$\int_0^{1/2} \phi(x) dx + \Phi = 1$$

La probabilité d'obtenir l'échantillon F devient ;

$$\int_0^{1/2} \phi(x) \cdot P(F; x) dx + \Phi \cdot P(F; 1/2)$$

et la distribution "a posteriori" de x, $\phi^*(x)dx$, sachant que l'échantillon F a été observé, est, pour $0 \leq x \leq 1/2$:

$$\phi^*(x)dx = \frac{\phi(x) \cdot P(F; x)}{\int_0^{1/2} \phi(x) \cdot P(F; x) dx + \Phi \cdot P(F; 1/2)} = \frac{\phi(x) \cdot \lambda(F; x)}{\int_0^{1/2} \phi(x) \cdot \lambda(F; x) dx + \Phi}$$

pour $x = 1/2$, la probabilité Φ^* d'indépendance est :

$$\Phi^* = \frac{\Phi \cdot P(F; 1/2)}{\int_0^{1/2} \phi(x) \cdot P(F; x) dx + \Phi \cdot P(F; 1/2)} = \frac{\Phi}{\int_0^{1/2} \phi(x) \cdot \lambda(F; x) dx + \Phi}$$

Malheureusement, la distribution a priori de x n'est pas connue.

Il n'est donc pas possible de calculer de façon précise les probabilités Φ^* et $\phi^*(x)dx$. Cependant, par analogie avec la Drosophile, on peut prendre pour approximation :

$$\phi(x)dx = 2(1 - \Phi)dx$$

distribution uniforme de 0 à 1/2. L'Homme possédant 23 paires de chromosomes, $1 - \Phi$ est de l'ordre de 1/23 si l'on admet, ce qui est d'ailleurs faux, que tous les chromosomes ont la même longueur. (Si l'on tient compte des différences de longueur, on obtient :

$$\Phi = \frac{\sum L^2}{(\sum L)^2} = \frac{\bar{L}^2}{n(\bar{L})^2}$$

Ceci donne :

$$\phi^*(x)dx \approx \frac{2\lambda(F; x)}{22 + \Lambda(F)} \approx \frac{2\lambda(F; x)}{\Lambda(F)}$$

si $\Lambda(F)$ est grand.

$$\text{et : } \Phi^* \approx \frac{22}{22 + \Lambda(F)} \approx \frac{22}{\Lambda(F)}$$

si $\Lambda(F)$ est grand. Nous avons donc obtenu, pour le calcul de ces probabilités, des approximations ne dépendant que de $\lambda(F; x)$ et $\Lambda(F)$.

Celles-ci seront satisfaisantes si le maximum de la courbe représentative de $\lambda(F; x)$ est très net. L'expression de Φ^* , probabilité d'indépendance, montre qu'une valeur élevée de $\Lambda(F)$ indique une forte probabilité d'existence d'une liaison entre les deux loci.

Cette méthode permet donc une estimation, assez grossière, du pourcentage de recombinaisons, et le calcul de la probabilité de la liaison, lorsque $\lambda(F; x)$ et $\Lambda(F)$ sont connus.

b) - Méthode de l'intervalle de confiance.

Si ξ est la vraie valeur du pourcentage de recombinaisons entre les deux loci, la probabilité d'obtenir l'échantillon F est :

$$P(F; \xi)$$

et pour toute valeur de x on a :

$$E \left[\frac{\lambda(F; x)}{\lambda(F; \xi)} \right] = \sum \frac{\lambda(F; x)}{\lambda(F; \xi)} P(F; \xi) = \sum P(F; x) = 1$$

En sommant par rapport à x sur l'intervalle (0, 1/2) il vient :

$$E \left[\frac{\Lambda(F)}{\lambda(F; \xi)} \right] = E \left[\frac{2}{\lambda(F; \xi)} \int_0^{1/2} \lambda(F; x) dx \right] = 2 \int_0^{1/2} E \left[\frac{\lambda(F; x)}{\lambda(F; \xi)} \right] dx = 2 \int_0^{1/2} dx = 1$$

S'il y a indépendance des deux loci, $\xi = 1/2$, et, $\lambda(F; \xi) = 1$ donc $E[\Lambda(F)] = 1$. Pour toute valeur de x comprise dans l'intervalle $0 \leq x \leq 1/2$, $\Lambda(F)$ est une variable aléatoire positive. Il résulte de l'inégalité de MARKOFF que, en l'absence de liaison, pour tout nombre positif $K > 1$, on a :

$$Pr \left\{ \Lambda(F) > K \right\} < \frac{E[\Lambda(F)]}{K} = \frac{1}{K}$$

Prenons $K = 10$. Ceci signifie que l'on a moins de 10 chances sur 100 d'observer $\Lambda(F) > 10$ s'il y a indépendance.

On a de même :

$$Pr \left\{ \frac{\Lambda(F)}{\lambda(F; \xi)} > K \right\} < \frac{1}{K}$$

ce qui équivaut à :

$$Pr \left\{ \lambda(F; \xi) > \frac{\Lambda(F)}{K} \right\} > 1 - \frac{1}{K}$$

c'est-à-dire que l'on a plus de 90 chances sur 100 que ξ , vraie valeur du pourcentage de recombinaisons soit compris dans l'intervalle des x vérifiant :

$$\lambda(F; x) > \frac{\Lambda(F)}{10}$$

Cette méthode permet donc de déterminer un "intervalle de confiance" pour l'estimation du pourcentage de recombinaisons. Elle s'applique quelle que soit la taille de l'échantillon observé et, par conséquent, à chaque famille F_i considérée isolément. Il en résulte que :

$$E \left[\frac{\Lambda(F_i)}{\lambda(F_i; \xi)} \right] = 1$$

et, si les familles sont statistiquement indépendantes :

$$E \left[\frac{\Lambda(F_1) \cdot \Lambda(F_2) \dots \Lambda(F_n)}{\lambda(F_1; \xi) \cdot \lambda(F_2; \xi) \dots \lambda(F_n; \xi)} \right] = E \left[\prod_{i=1}^n \frac{\Lambda(F_i)}{\lambda(F_i; \xi)} \right] = 1$$

Nous avons donc au moins 90 chances sur 100 que ξ appartienne à l'en-

-semble des valeurs de x telles que :

$$\lambda(F; x) > \frac{\prod_{i=1}^n \lambda(F_i)}{10}$$

Test d'homogénéité.

Nous avons vu que $\Lambda(F)$ prend des valeurs élevées lorsqu'il y a liaison, et a une espérance mathématique égale à 1 lorsqu'il y a indépendance. $\Lambda(F_i)$ a une signification analogue pour la famille F_i prise isolément. Par conséquent le produit $\prod_{i=1}^n \Lambda(F_i)$ deviendra grand lorsqu'il y aura liaison dans toutes les familles, ou dans certaines d'entre elles, même si le pourcentage de recombinaisons varie selon les familles. En revanche, si le pourcentage $\%$ de recombinaisons est identique dans toutes les familles, la plupart des courbes représentatives des $\lambda(F_i; x)$ doivent passer par un maximum pour une valeur de x proche de $\frac{1}{2}$. La valeur de $\Lambda(F)$ est alors très supérieure au produit des $\Lambda(F_i)$. S'il n'en est pas ainsi, c'est-à-dire si $\Lambda(F_i) > \Lambda(F)$, on peut penser que les différentes familles n'ont pas toutes le même pourcentage de recombinaisons, autrement dit, que l'échantillon étudié est hétérogène. Cette hétérogénéité peut être due, nous l'avons vu, à la présence d'accidents chromosomiques, tels que des inversions, dans certaines familles, ou bien encore, le caractère principal peut être déterminé par deux ou plusieurs mutations à des loci différents, un seul d'entre eux étant lié au locus marqueur considéré. Il en est ainsi, par exemple, pour la liaison elliptocytose-Rh qui a permis de distinguer deux formes d'elliptocytose. L'étude des liaisons entre loci permet donc, dans ces cas favorables, de distinguer plusieurs variétés, dues à des mutations différentes, d'un même tableau clinique.

Mutations.

Il y a lieu parfois, de tenir compte de la possibilité de mutation à l'un des loci étudiés. HALDANE (1935), a montré par exemple, que le gène de l'hémophilie disparaît, en moyenne, après trois générations. Il faut donc s'attendre, en compensation, à voir apparaître le gène par mutation dans un certain nombre de familles. Dans la grande majorité des familles étudiées par HALDANE et SMITH (1947), la mutation au locus de l'hémophilie est survenue avant qu'il soit possible d'obtenir une information sur le locus de la cécité aux couleurs. Le seul effet de cette mutation sur la probabilité $P(F_i; x)$ de la famille F_i considérée, est de la multiplier par un facteur indépendant de x qui s'élimine par conséquent lorsque l'on forme le rapport des probabilités. La mutation ne modifie donc pas la valeur de $\lambda(F_i; x)$. En revan-

60

-che, dans quelques familles, telles que la famille B de HOOGVLIET, la possibilité d'une mutation, introduit un ou plusieurs facteurs dépendant de x , que l'on ne négligerait pas sans une erreur appréciable.

- Distinction entre les sexes des parents dans les études de liaisons.

On connaît, en Génétique Animale, diverses espèces dans lesquelles la valeur de x , pourcentage de recombinaisons, diffère chez le mâle et chez la femelle. Il est donc intéressant de disposer d'un test statistique permettant de calculer séparément les valeurs de x des deux sexes (C.A.B.SMITH, 1954). Il faut cependant noter qu'un test supposant des valeurs de x égales chez l'homme et chez la femme, est parfaitement correct. En effet, le score " λ ", calculé selon la méthode de HALDANE et SMITH (1947), sans distinction des sexes, est simplement la somme des scores paternel et maternel. Enfin, une valeur de x différenciant significativement de 0,5 dans l'un ou l'autre sexe, suffit à établir l'existence d'une liaison.

a)- Cas des familles dont les génotypes des parents sont connus aux deux loci, excepté en ce qui concerne la position en phase ou en répulsion.

Nous appellerons G, g, G', g' , le système d'allèles déterminant le caractère principal; celui qui contrôle le caractère marqueur sera désigné par T, t, T', t' .

Considérons une famille f : soit $GgTt$ le génotype de la mère, $G'g'T't'$ celui du père. L'utilisation des majuscules ne signifie pas nécessairement que le gène G est dominant sur g ou différent de g mais permet simplement de distinguer les deux gènes d'un même individu. En ce qui concerne le caractère marqueur, nous pouvons représenter ce mariage par : $Tt \times T't'$. Chaque enfant reçoit de sa mère, soit le gène T , soit le gène t . Nous répartissons les enfants de ce mariage en cinq classes, selon leur probabilité d'avoir reçu le gène T de leur mère; à chaque classe, nous faisons correspondre une certaine valeur e .

-1/- enfants "T-certains".

Ce sont ceux pour lesquels on peut être sûr que le gène hérité de leur mère est T . C'est le cas, par exemple, de tous les enfants de groupe sanguin A, issus d'un mariage $Tt \times T't' = AB \times OO$.

Soient P la probabilité que le gène marqueur venant de la mère soit

T et Q la probabilité, complémentaire, que ce gène soit t. La valeur e est définie par :

$$e = P - Q = 2P - 1$$

Pour la classe des enfants T-certains, nous avons :

$$P = 1, \quad Q = 0, \quad e = 1$$

-2/- enfants "T-impossibles".

Ce sont ceux dont nous sommes sûrs que le gène transmis par leur mère n'est pas T. Il en est ainsi pour les enfants de groupe sanguin B issus du mariage cité au paragraphe précédent. Dans ce cas,

$$P = 0, \quad Q = 1, \quad e = -1$$

-3/- enfants "T-indifférents".

Lorsqu'il y a autant de chances pour que le gène transmis par la mère soit T que t, on dit que l'enfant est t-indifférent. C'est le cas, par exemple, des enfants issus d'un mariage $Tt \times T't' = OO \times AB$; tous les enfants possèdent un gène O venant de leur mère sans qu'il soit possible de dire si c'est ou non celui qui a été arbitrairement désigné par T. Ici :

$$P = Q = 1/2, \quad e = 0$$

-4/- enfants "T-probables".

Considérons cette fois le mariage $Tt \times T't' = BO \times BO$. Un enfant de groupe B issu de ce mariage, a autant de chances d'être de génotype BB que BO, B venant de sa mère, ou OB, O venant de sa mère. Ainsi, dans deux cas sur trois, cet enfant hérite le gène $T = B$ de sa mère.

On peut donc dire qu'il est "T-probable", et :

$$P = 2/3, \quad Q = 1/3, \quad e = 1$$

-5/- enfants "T-improbables".

Inversement, si le même mariage est désigné cette fois par : $Tt \times T't' = OB \times OB$, c'est le gène $T = O$ qui aura la plus faible probabilité d'être transmis à l'enfant de groupe B par sa mère. Il est dit "T-improbable" et :

$$P = 1/3, \quad Q = 2/3, \quad e = -1/3$$

Les quatre génotypes, également probables, pour les enfants issus du mariage $Tt \times T't'$ sont : TT' , que nous désignerons plus brièvement par T_1 , $Tt' = \theta_1$, $tT' = \theta'_1$ et $tt' = t_1$. Dans certains mariages, ces génotypes correspondent à quatre phénotypes différents. Par exemple, $Tt \times T't' = AO \times BO$:

$$T_1 = TT' = AB = [AB]$$

$$\theta_1 = Tt' = AO = [AO]$$

$$t_1 = tt' = OO = [O]$$

$$\theta'_1 = tT' = OB = [B]$$

Dans d'autres cas, il est impossible de distinguer les quatre génotypes d'après les phénotypes, soit parce que l'un des allèles est récessif, soit parce qu'un même gène est présent à plusieurs exemplaires. Si l'on ne peut pas distinguer $T_1 = TT'$ de $\Theta_1 = Tt'$, le phénotype correspondant est désigné par T_2 .

Tableau I

Classification des différentes possibilités des phénotypes.

$$\begin{aligned}
 T_1 &= TT' \\
 \Theta_1 &= Tt' \\
 \Theta'_1 &= tT' \\
 t_1 &= tt' \\
 T_2 &= T_1 \text{ ou } \Theta_1 = TT' \text{ ou } Tt' \\
 t_2 &= \Theta'_1 \text{ ou } t_1 = tT' \text{ ou } tt' \\
 T'_2 &= T_1 \text{ ou } \Theta'_1 = TT' \text{ ou } tT' \\
 t'_2 &= \Theta_1 \text{ ou } t_1 = Tt' \text{ ou } tt' \\
 \textcircled{H}_2 &= \Theta_1 \text{ ou } \Theta'_1 = Tt' \text{ ou } tT' \\
 \textcircled{H}^*_2 &= T_1 \text{ ou } t_1 = TT' \text{ ou } tt' \\
 T_3 &= T_1 \text{ ou } \Theta_1 \text{ ou } \Theta'_1 = TT' \text{ ou } Tt' \text{ ou } tT' \\
 \Theta'_3 &= T_1 \text{ ou } \Theta'_1 \text{ ou } t_1 = TT' \text{ ou } tT' \text{ ou } tt' \\
 \Theta_3 &= T_1 \text{ ou } \Theta_1 \text{ ou } t_1 = TT' \text{ ou } Tt' \text{ ou } tt' \\
 t_3 &= \Theta_1 \text{ ou } \Theta'_1 \text{ ou } t_1 = Tt' \text{ ou } tT' \text{ ou } tt' \\
 T_4 &= T_1 \text{ ou } \Theta_1 \text{ ou } \Theta'_1 \text{ ou } t_1 = TT' \text{ ou } Tt' \text{ ou } tT' \text{ ou } tt'
 \end{aligned}$$

Tableau II

Classification phénotypique de l'enfant selon la probabilité P que le gène reçu de sa mère soit T . Valeurs correspondantes de e .

classe	P	$Q = 1 - P$	$e = P - Q$	phénotypes possibles
1/- T-certains	1	0	1	T_1, Θ_1, T_2
2/- T-possibles	2/3	1/3	1/3	T_3, Θ_3
3/- T-indifférents	1/2	1/2	0	$T'_2, t'_2, T_4, \textcircled{H}_2, \textcircled{H}^*_2$
4/- T-improbables	1/3	2/3	-1/3	Θ'_3, t_3
5/- T-impossibles	0	1	-1	Θ_1, t_1, t_2

On introduit également une classification pour le caractère principal I = G-certain, II = G-probable, III = G-indifférent, IV = G-improbable et V = G-impossible selon que la probabilité Φ que le gène transmis par la mère à l'enfant soit G, est égale à 1, 2/3, 1/2, 1/3 ou 0.

Posons $\Psi = 1 - \Phi$ et $\Xi = \Phi - \Psi = 1, 1/3, 0, -1/3$ ou -1 selon que l'enfant appartient à la classe I, II, III, IV ou V. Les différents phénotypes possibles, définis de la même façon que pour le caractère marqueur sont :

$$G_1, \gamma_1, \gamma_1', g_1, G_2, g_2, G_2', g_2', \Gamma_2, \Gamma_2', G_3, \gamma_3, \gamma_3', g_3, \text{ et } G_4.$$

Supposons que la famille F comporte n enfants désignés par $C_1, C_2, \dots, C_n$. A chaque enfant C_i , correspond une valeur $e = e(C_i)$ et une valeur $\Xi = \Xi(C_i)$. Soit $\alpha = \alpha(C_i) = \Xi(C_i) \cdot e(C_i)$. Le score $\lambda(F)$ de liaison maternelle est défini par :

$$\lambda(F) = \sum_{i < j} \alpha(C_i) \cdot \alpha(C_j)$$

(chaque paire $C_i C_j$ n'est comptée qu'une fois). Cette expression peut également s'écrire sous la forme, plus facile à utiliser :

$$\lambda(F) = 1/2 \left(\sum_1^n \alpha \right)^2 - 1/2 \sum (\alpha^2)$$

Pour les exemples de calcul de $\lambda(F)$, dans les différents types de mariages, nous renvoyons le lecteur au travail original de C.A.B. SMITH. Nous noterons seulement que dans cette méthode de calcul du score λ , seuls les enfants dont les phénotypes sont connus pour les deux caractères, peuvent apporter une information.

On définit de même $\lambda'(F)$, score de liaison paternelle, obtenu à partir des valeurs Ξ' , e' , et α' calculées de la même façon que Ξ , e et α , mais cette fois par rapport au père. Le score total de la famille, λ^+ , sans distinction du sexe, égal au λ du paragraphe précédent, est la somme des scores paternel et maternel.

$$\lambda^+(F) = \lambda(F) + \lambda'(F)$$

Il faut remarquer que la distribution de $\lambda(F)$ risque d'être biaisée si la sélection des familles à étudier se fait en fonction de la présence d'enfants tarés, et non pas en fonction du génotype des parents déterminé directement sur un échantillon de couples tiré au hasard. Les génotypes parentaux peuvent également, sans fausser la distribution, être déterminés grâce à des informations fournies par les phénotypes de sujets apparentés, autres que les enfants qui seront utilisés pour le calcul du score. Si l'on est obligé de tenir compte du génotype de certains enfants pour déterminer celui des parents, ces enfants doivent être omis dans le calcul de λ .

- Propriétés du score " λ ".

a) - calcul de l'espérance mathématique de λ : $E[\lambda(F)] = 0$.

Lemme 1 : l'espérance mathématique de $e(C_i)$, pour tout enfant C_i , est nulle.

Soient $T_1, T_2, \dots$ etc, les différents phénotypes possibles pour les enfants $C_1, C_2, \dots, C_n$, d'un mariage donné F , et $\pi(T_1), \pi(T_2) \dots$ etc, leurs probabilités respectives. T_i est le phénotype de l'enfant C_i , $\pi(T_i)$, la probabilité de ce phénotype.

$$e(T_i) = P(T_i) - Q(T_i)$$

où $P(T_i)$ est la probabilité qu'un enfant de phénotype T_i ait reçu le gène T de sa mère, et $Q(T_i)$, la probabilité que ce soit t qu'il ait reçu de celle-ci.

$$E[e(C_i)] = \sum \pi(T_i) \cdot e(C_i) = \sum \pi(T_i) \cdot P(T_i) - \sum \pi(T_i) \cdot Q(T_i)$$

La sommation s'étend à tous les phénotypes T_i . $\sum \pi(T_i) \cdot P(T_i)$ est la probabilité pour qu'un enfant, choisi au hasard, ait reçu son gène T de sa mère, probabilité égale à $1/2$. On a de même, $\sum \pi(T_i) \cdot Q(T_i) = 1/2$. Par conséquent :

$$E[e(C_i)] = 0$$

Les phénotypes des différents enfants sont statistiquement indépendants. Pour tout couple $C_i C_j$, nous avons donc :

$$E[e(C_i) \cdot e(C_j)] = E[e(C_i)] \cdot E[e(C_j)] = 0$$

Dans l'hypothèse nulle, de l'indépendance, le caractère principal et le caractère marqueur sont transmis indépendamment; la valeur de $\lambda(C_i)$ est donc indépendante de celle de $e(C_i)$. Par conséquent :

$$\begin{aligned} E[\alpha(C_i) \cdot \alpha(C_j)] &= E[\varepsilon(C_i) \cdot \varepsilon(C_j) \cdot e(C_i) \cdot e(C_j)] \\ &= E[\varepsilon(C_i) \cdot \varepsilon(C_j)] \cdot E[e(C_i) \cdot e(C_j)] = 0 \end{aligned}$$

Ainsi, dans l'hypothèse nulle, l'espérance mathématique du score maternel est nulle.

Nous avons de la même façon : $E[\lambda'(F)] = 0$

Il faut donc, pour pouvoir affirmer qu'il existe une liaison, prouver que $E[\lambda]$ est significativement plus grande que 0. Pour ce faire, nous calculons l'erreur standard de $\lambda(F)$, donnée par l'expression :

$$\Delta = 1 / \sqrt{d^2 \lambda / dx^2}$$

Si $\lambda(F)$ est supérieur à 1,6s, nous admettons en effet que λ est distribué normalement, nous considérons le résultat comme significatif avec une probabilité de 5%, ce qui veut dire que, si les loci sont indépendants, nous n'avons que 5 chances sur 100 d'obtenir une valeur de $\lambda(F)$ qui soit supérieure à 1,6s. Si $\lambda(F)$ est supérieur à 2,33s, le résultat est significatif avec une sécurité de 1%. Ceci suppose

que $\lambda(F)$ est distribué normalement, ce qui est vrai pour les grands échantillons mais peut être très éloigné de la réalité pour les petits (HALDANE, 1947).

b) - calcul de la variance dans l'hypothèse nulle.

Si l'on admet que les contributions $\alpha(C_i) \cdot \alpha(C_j)$ à $\lambda(F)$, des différentes paires de germains, sont statistiquement indépendantes, la variance de $\lambda(F)$ est égale à la somme des variances des différentes paires ij . Nous pouvons donc écrire :

$$\begin{aligned} \text{var } \lambda(F) &= \sum \text{var } [\alpha(C_i) \cdot \alpha(C_j)] \\ \text{var } [\alpha(C_i) \cdot \alpha(C_j)] &= E [\alpha(C_i) \cdot \alpha(C_j)]^2 - \underbrace{(E [\alpha(C_i) \cdot \alpha(C_j)])^2}_{= 0} \\ &= E [\varepsilon(C_i) \cdot \varepsilon(C_j) \cdot e(C_i) \cdot e(C_j)]^2 \\ &= E [\varepsilon(C_i)]^2 \cdot E [\varepsilon(C_j)]^2 \cdot E [e(C_i)]^2 \cdot E [e(C_j)]^2 \end{aligned}$$

Le terme $E [\alpha(C_i) \cdot \alpha(C_j)]^2$ est nul, comme nous venons de le voir. $E [e(C_i)]^2$ est l'espérance mathématique de $[e(C_i)]^2$ et est donc indépendante de l'enfant C_i choisi. Posons :

$$E [e(C_i)]^2 = E [e(C_j)]^2 = \sqrt{d}$$

Il vient alors :

$$\text{var } [\alpha(C_i) \cdot \alpha(C_j)] = d \cdot \sum E [\varepsilon(C_i)]^2 \cdot E [\varepsilon(C_j)]^2$$

d'où :

$$\text{var } [\lambda(F)] = d \cdot \sum E [\varepsilon(C_i)]^2 \cdot E [\varepsilon(C_j)]^2$$

De même, si l'on pose :

$$\sqrt{d'} = E [e(C_i) \cdot e'(C_i)]$$

$$\sqrt{d''} = E [e'(C_i)]^2 = E [e'(C_j)]^2$$

il vient :

$$\text{cov } [\lambda, \lambda'] = d' \cdot \sum E [\varepsilon(C_i) \cdot \varepsilon'(C_i)] \cdot E [\varepsilon(C_j) \cdot \varepsilon'(C_j)]$$

$$\text{var } [\lambda'(F)] = d'' \cdot \sum E [e'(C_i)]^2 \cdot E [e'(C_j)]^2$$

$E [e(C_i)]^2$, peut être calculé directement dans tous les cas particuliers. Pour pouvoir donner une règle générale de calcul de d , d' , d'' , il est nécessaire de classer les différents types possibles de ségrégation pour le caractère marqueur. Selon le type de mariage, on obtient des enfants pouvant avoir, quatre phénotypes distincts, classe 1111, deux phénotypes dans le rapport 3/1, classe 31, etc..., un seul phénotype, classe 4. Selon le type de famille, les valeurs de d , d' , et d'' sont :

type de voisement pour le caractère marqueur	d	d'	d''
[1, 1, 1, 1]	1	0	1
[2, 2]	1	0	0
[2, 2']	0	0	1
[2, 1, 1]	1	0	1/4
[2, 1, 1']	1/4	0	1
[1, 2, 1]	1/4	1/4	1/4
[3, 1]	1/9	1/9	1/9
[2, 2*] ou [4]	0	0	0

Si nous supposons maintenant que le repérage des familles selon le caractère principal se fait par les parents, nous avons, de la même façon, pour le caractère principal :

$$E[E(C_i)]^2 = \sqrt{\delta}, \quad E[E(C_i) \cdot E'(C_i)] = \sqrt{\delta'}, \quad E[E'(C_i)]^2 = \sqrt{\delta''}$$

δ , δ' , et δ'' jouent, pour le caractère principal, le même rôle que d , d' et d'' pour le caractère marqueur. Il vient alors :

$$\text{var } \lambda(F) = d \Sigma \delta$$

la sommation porte sur toutes les paires $C_i C_j$. S'il y a n frères et soeurs, il est possible de former $\frac{n(n-1)}{2}$ paires, par conséquent :

$$\text{var } \lambda(F) = 1/2 n(n-1) d \delta$$

$$\text{covar}[\lambda(F), \lambda'(F)] = 1/2 n(n-1) d' \delta'$$

$$\text{var } \lambda'(F) = 1/2 n(n-1) d'' \delta''$$

Considérons maintenant un échantillon de k familles, $F_1, \dots, F_i, \dots, F_k$, pour lesquelles les scores $\lambda(F_i)$, $\lambda'(F_i)$, leurs variances et covariances ont été calculés selon la méthode exposée ci-dessus. Les score, variance et covariance totaux, s'obtiennent par sommation sur toutes les familles puisque celles-ci sont indépendantes.

$$\lambda = \sum_{i=1}^k [\lambda(F_i)] \quad , \quad \lambda' = \sum_{i=1}^k [\lambda'(F_i)]$$

$$v = \text{var } \lambda = \sum_{i=1}^k \text{var } \lambda(F_i)$$

$$v' = \text{cov}(\lambda, \lambda') = \sum_{i=1}^k \text{cov}[\lambda(F_i), \lambda'(F_i)]$$

$$v'' = \text{var } \lambda' = \sum_{i=1}^k \text{var } \lambda'(F_i)$$

Si les deux loci sont liés, λ est supérieur à une certaine valeur dépendant de son erreur standard, $s = \sqrt{v}$, et de la précision demandée λ sera significatif à 5% si $\lambda > 1,6 s$, si l'on peut admettre que la

distribution de λ suit la loi normale. Il en est de même pour λ' . Nous avons donc un moyen de tester la liaison séparément pour chaque sexe. Pour appliquer ce test sans considération de sexe, nous calculons :

$$\lambda^+ = \lambda + \lambda'$$

$$v^+ = v + 2v' + v''$$

λ^+ est significatif à 5% si $\lambda^+ > 1,6 \sqrt{v + 2v' + v''}$

c) - échantillonnage faussé par la sélection sur le caractère principal

En pratique, les familles sont toujours sélectionnées à partir des enfants tarés pour le caractère principal; le génotype des parents, au locus principal, est déterminé soit directement, soit à partir des enfants. Ces familles sont ensuite testées pour différents caractères marqueurs. Ce mode de sélection des familles introduit une erreur difficile à apprécier exactement, mais qui peut être importante.

Ce biais s'introduit seulement par l'intermédiaire du caractère principal, le caractère marqueur est, en effet, distribué au hasard, tout au moins dans l'hypothèse nulle. Différentes méthodes de correction de cette erreur ont été proposées, en particulier par FISHER et par FINNEY; la plus simple est de considérer le caractère principal comme fixé, et d'étudier seulement la ségrégation du caractère marqueur. Nous allons donc choisir maintenant, dans notre échantillon, les familles pour lesquelles le génotype des parents, au locus marqueur, est connu. $E(\lambda)$ est encore nulle, mais puisque le caractère principal est considéré comme fixé, $E(C_i)$ et $E'(C_i)$ sont aussi fixés et l'on a :

$$E[E(C_i)]^2 = [E(C_i)]^2$$

Il vient donc :

$$\text{var } \lambda(F) = d \cdot \sum [E(C_i)]^2 \cdot [E(C_j)]^2$$

$$\text{cov}[\lambda(F), \lambda'(F)] = 1/2 d' [(\sum \epsilon \epsilon')^2 - \sum (\epsilon \epsilon')^2]$$

$$\text{var } \lambda'(F) = d'' \cdot \sum [E'(C_i)]^2 \cdot [E'(C_j)]^2$$

La sommation, comme plus haut, porte sur toutes les paires $C_i C_j$. Ceci peut encore s'écrire :

$$\text{var } \lambda(F) = 1/2 d [(\sum \epsilon^2)^2 - \sum \epsilon^4]$$

$$\text{cov}[\lambda(F), \lambda'(F)] = 1/2 d' [(\sum \epsilon \epsilon')^2 - \sum (\epsilon \epsilon')^2]$$

$$\text{var } \lambda'(F) = 1/2 d'' [(\sum \epsilon'^2)^2 - \sum \epsilon'^4]$$

en sommant par rapport aux enfants pris individuellement.

b) - Cas des parents de génotype inconnu, ou non testés directement.

Presque tous les échantillons de données utilisés pour les études de liaisons en Génétique Humaine, comportent au moins quelques familles où les génotypes des parents sont partiellement inconnus. Pour chacune d'entre elles, il existe un certain nombre de types de mariages compatibles avec les phénotypes observés chez les différents membres de la famille. Soient $M_1, M_2, \text{etc.}$ ces diverses possibilités. Chaque mariage M_i a une probabilité $\pi(M_i)$ d'être observé dans la population générale. $\pi(M_i)$ est calculé à partir de la fréquence des gènes considérés dans la population. Pour chaque croisement M_i , on calcule, d'après les lois de MENDEL, la probabilité $\pi(F/M_i)$ de rencontrer une famille telle que F dans l'hypothèse nulle d'indépendance, et, également, les scores $\lambda(F/M_i)$ et $\lambda'(F/M_i)$ dans l'hypothèse alternative de liaison des deux loci. Les scores réels $\lambda(F)$ et $\lambda'(F)$ sont les moyennes, pondérées par le produit de probabilités $\pi(M_i) \cdot \pi(F/M_i)$, de $\lambda(F/M_i)$ et $\lambda'(F/M_i)$ c'est-à-dire :

$$\lambda(F) = \frac{\sum [\pi(M_i) \cdot \pi(F/M_i) \cdot \lambda(F/M_i)]}{\sum [\pi(M_i) \cdot \pi(F/M_i)]}$$

$$\lambda'(F) = \frac{\sum [\pi(M_i) \cdot \pi(F/M_i) \cdot \lambda'(F/M_i)]}{\sum [\pi(M_i) \cdot \pi(F/M_i)]}$$

En posant $\pi(M_i/F)$ = probabilité a posteriori que le mariage soit en fait M_i , sachant que la famille F a été observée, soit :

$$\pi(M_i/F) = \frac{\pi(M_i) \cdot \pi(F/M_i)}{\sum_j [\pi(M_j) \cdot \pi(F/M_j)]}$$

nous avons alors :

$$\lambda(F) = \sum \pi(M_i/F) \cdot \lambda(F/M_i)$$

$$\lambda'(F) = \sum \pi(M_i/F) \cdot \lambda'(F/M_i)$$

Malheureusement, le calcul de la variance du score λ obtenu de cette façon est souvent fort complexe (FINNEY). Cependant, le score λ ainsi calculé pour une famille où le génotype des parents n'est pas connu, peut être interprété comme l'espérance mathématique du score que l'on obtiendrait pour cette même famille s'il était possible de déterminer le type de mariage dont il s'agit. Lorsque le type de mariage M_i d'une famille F est connu, la variance est :

$$v(F/M_i) = \text{var } \lambda(F/M_i)$$

L'espérance mathématique de la variance de cette famille est alors :

$$V_1 = \sum \pi(M_i/F) \cdot v(F/M_i)$$

mais il faut noter que la connaissance du type de mariage, M_i , de

chaque famille, remplace le score $\lambda(F)$ par un score $\lambda(F/M_i)$ qui dépend de M_i . Ceci introduit un terme supplémentaire, V_2 , dans l'expression de la variande :

$$V_2 = \sum_i \pi(M_i/F) [\lambda(F/M_i)]^2 - [\lambda(F)]^2$$

La variance V de $\lambda(F)$, lorsque les types de mariages ne sont pas connus, est la différence $V_1 - V_2$. De la même façon :

$$v'(F/M_i) = \text{cov}[\lambda(F/M_i), \lambda'(F/M_i)]$$

$$v''(F/M_i) = \text{var}[\lambda'(F/M_i)]$$

$$V_1' = \sum_i \pi(M_i/F) \cdot v'(F/M_i)$$

$$V_1'' = \sum_i \pi(M_i/F) \cdot v''(F/M_i)$$

$$V_2' = \sum_i \pi(M_i/F) \cdot \lambda(F/M_i) \cdot \lambda'(F/M_i) - \lambda(F) \cdot \lambda'(F)$$

$$V_2'' = \sum_i \pi(M_i/F) [\lambda'(F/M_i)]^2 - [\lambda'(F)]^2$$

$$V' = V_1' - V_2' \quad V'' = V_1'' - V_2''$$

La méthode de calcul qui vient d'être décrite pour le cas où le génotype des parents est inconnu, est d'une utilisation extrêmement laborieuse dès qu'il y a plusieurs types de mariages possibles à envisager. Ce travail pénible apporte, en fin de compte, un supplément d'information très minime. Presque toute l'information utilisée est fournie par les familles où le génotype des parents est connu, et le calcul du score de ces familles est relativement facile. Néanmoins, dans un certain nombre de cas, il est possible de déterminer avec certitude le génotype des parents à condition de tenir compte du génotype de certains enfants utilisés également dans le calcul du score ce qui, nous l'avons vu, introduit une erreur. FINNEY (1940), a proposé une méthode de correction du score, applicable, en principe à tous les types de familles, et plus précise que la méthode précédente. Mais son utilisation peut être extrêmement pénible dans certains cas.

Supposons que les parents d'une famille donnée, F , forment un couple de type M_x , et qu'ils puissent donner naissance à différentes possibilités $f_1, f_2, \dots, f_k$ de familles de n enfants. Le type F_i a une probabilité π_i dans l'hypothèse nulle. Soient $\lambda_i, \lambda'_i, v = \text{var} \lambda_i, v' = \text{cov}(\lambda_i, \lambda'_i)$ et $v'' = \text{var} \lambda'_i$, les scores, variances et covariances de la famille F . Certaines de ces possibilités, disons f_1, f_2, f_3 , par exemple, ne permettent pas la détermination du type de mariage M_x . Ils doivent être rejetés de l'échantillon. Les diverses possibilités restantes, $f_4, f_5, \dots, f_k$ sont observées en proportions $\pi_4, \pi_5, \dots, \pi_k$.

et leurs probabilités respectives sont par conséquent :

$$\pi_4^* = \frac{\pi_4}{\pi_4 + \pi_5 + \dots + \pi_k} \quad \pi_5^* = \frac{\pi_5}{\pi_4 + \pi_5 + \dots + \pi_k} \quad \text{etc...}$$

La valeur moyenne de λ est donc :

$$\bar{\lambda} = \pi_4^* \lambda_4 + \pi_5^* \lambda_5 + \dots + \pi_k^* \lambda_k = \frac{\pi_4 \lambda_4 + \pi_5 \lambda_5 + \dots + \pi_k \lambda_k}{\pi_4 + \pi_5 + \dots + \pi_k}$$

Mais $\pi_1 + \pi_2 + \pi_3 + \pi_4 + \dots + \pi_k = 1$ et $\pi_1 \lambda_1 + \dots + \pi_k \lambda_k = 0$

Nous pouvons donc écrire :

$$\bar{\lambda} = \frac{-\pi_1 \lambda_1 - \pi_2 \lambda_2 - \pi_3 \lambda_3}{1 - \pi_1 - \pi_2 - \pi_3}$$

et l'espérance mathématique du score corrigé, $\lambda - \bar{\lambda}$ est bien nulle.
 $\bar{\lambda}$ est la correction du score.

La correction pour la variance s'obtient de la même façon :

$$\text{var} (\lambda - \bar{\lambda}) = \frac{\nu - \pi_1 \lambda_1^2 - \pi_2 \lambda_2^2 - \pi_3 \lambda_3^2}{1 - \pi_1 - \pi_2 - \pi_3} - \bar{\lambda}^2$$

$$\text{cov} (\lambda - \bar{\lambda})(\lambda' - \bar{\lambda}') = \frac{\nu' - \pi_1 \lambda_1 \lambda'_1 - \pi_2 \lambda_2 \lambda'_2 - \pi_3 \lambda_3 \lambda'_3}{1 - \pi_1 - \pi_2 - \pi_3} - \bar{\lambda} \bar{\lambda}'$$

$$\text{var} (\lambda' - \bar{\lambda}') = \frac{\nu'' - \pi_1 \lambda_1'^2 - \pi_2 \lambda_2'^2 - \pi_3 \lambda_3'^2}{1 - \pi_1 - \pi_2 - \pi_3} - \bar{\lambda}'^2$$

- METHODE SEQUENTIELLE DE N.E.MORTON.

Les données permettant l'étude des liaisons s'accumulent, en Génétique Humaine, sous la forme d'une collection d'échantillons de une ou plusieurs familles, dont chacun est très petit en comparaison de la quantité d'observations nécessaire pour pouvoir affirmer ou rejeter l'existence d'une liaison, même de valeur moyenne. Chaque nouvelle famille observée apporte une information supplémentaire. Il est donc logique d'appliquer à ces recherches (MORTON, 1955), les méthodes de l'analyse séquentielle (WALD, 1947).

Le test séquentiel est simple et efficace. Des tables de "lod-scores" (lod=log odds= log probability ratio), établies par MORTON et complétées par S.MAYNARD-SMITH, L.S.PENROSE et C.A.B.SMITH, réduisent au maximum les calculs. De plus, les systèmes de scores non séquentiels décrits par PENROSE, méthode des paires de germains, FISHER : scores-U FINNEY, BAILEY et SMITH : rapport de probabilités, permettent surtout de reconnaître l'existence d'une liaison, alors que les scores séquentiels permettent à la fois la détection et l'estimation des liaisons.

- Généralités sur les tests séquentiels.

Tester un échantillon F de n observations, c'est utiliser une règle spécifiant pour chaque échantillon possible de taille n, si l'hypothèse proposée, hypothèse nulle : les deux loci sont indépendants, peut être acceptée ou rejetée. Nous devons donc choisir entre deux hypothèses : H_0 , hypothèse nulle ou d'indépendance c'est-à-dire $x=1/2$, et H_1 , hypothèse alternative ou de liaison d'intensité x . A chaque nouvelle observation correspondent trois possibilités :

- a)- accepter H_0
- b)- rejeter H_0
- c)- continuer les observations

On admet, et il est démontré que le processus se terminera, c'est-à-dire qu'au bout d'un nombre fini d'essais, nous aboutirons soit au cas a) soit au cas b). Le nombre d'observations requis pour ce test est donc une variable aléatoire, puisqu'il dépend de l'issue de chaque expérience.

Soit F_n l'ensemble de tous les échantillons possibles de taille n. Il comprend trois parties : F_n^0 , ensemble de tous les échantillons de taille n pour lesquels H_0 est acceptée, F_n^1 , ensemble de ceux pour lesquels c'est H_1 qui est acceptée, $F_n^?$, ensemble des échantillons pour lesquels on ne peut admettre ni H_0 ni H_1 et qui obligent à conti-

-nuer les observations.

$$F_n = F_n^0 \cup F_n^1 \cup F_n^?$$

Si l'échantillon observé, de taille n , appartient :

- à F_n^0 , on accepte H_0
- à F_n^1 , on rejette H_0
- à $F_n^?$, on continue.

La partie ou sous-ensemble $F_n^?$ est appelée la "région critique". Le choix d'une méthode de test implique donc la connaissance de la région critique. Il y a, en général, une infinité de manières de la choisir; elles ne sont pas toutes aussi satisfaisantes, il faut donc donner quelques principes pour guider ce choix.

En acceptant ou en rejetant H_0 , nous pouvons commettre deux types d'erreur :

- 1/- en rejetant H_0 quand H_0 est vraie, c'est l'erreur dite de première espèce; nous appellerons α sa probabilité.
- 2/- en acceptant H_0 quand H_1 est vraie, c'est l'erreur de deuxième espèce de probabilité β .

Lorsque la région critique a été choisie, les probabilités α et β de commettre ces deux erreurs sont déterminées. Le meilleur test sera obtenu pour α et β petits. Or, si l'on se donne l'un des deux, l'autre est déterminé quand n est fixé. Si l'on considère toutes les régions critiques pour lesquelles α est fixé, la meilleure d'entre elles sera celle qui aura le plus faible β . α est la "taille" de la région critique, $1-\beta$ son "efficacité". Lorsque α et β sont fixés, le plus judicieusement possible, nous verrons comment un peu plus loin, il devient possible de déterminer $n(\alpha, \beta)$, nombre minimum d'observations nécessaires.

- Méthode de MORTON.

Soit x le pourcentage de recombinaisons entre les deux loci étudiés. L'échantillon de familles, F , est une variable aléatoire pouvant prendre les valeurs $f_1, f_2, \dots, f_n$, et telle que :

$$P(F = f_i) = p_i(x)$$

Dans l'hypothèse nulle, d'indépendance, H_0 , $x = 1/2$. Dans l'hypothèse alternative, de liaison, H_1 , $x = x_1$. Les probabilités d'obtenir l'échantillon $F = (f_1, f_2, \dots, f_n)$ sont :

- lorsque H_0 est vraie :

$$p_n^0 = p_1(1/2) \cdot p_2(1/2) \dots p_n(1/2)$$

- Lorsque H_1 est vraie :

$$p_n^1 = p_1(x_1) \cdot p_2(x_2) \dots p_n(x_n)$$

Le test séquentiel utilise la rapport de probabilités p_n^1 / p_n^0 et deux nombres positifs A et B, $A > 1$, $B < 1$, définis en fonction des erreurs de première et deuxième espèces α et β . Pour la commodité des calculs, on travaille sur les logarithmes :

$$\log \frac{p_n^1}{p_n^0} = \sum_{i=1}^n \log \frac{p_i(x_i)}{p_i(1/2)} = \sum_{i=1}^n Z_i = Z$$

Z est, par définition, le "lod-score" ou score logarithmique de l'échantillon F des n familles $f_1, f_2, \dots, f_n$. A chaque apport nouveau d'information, f_{j+1} , la somme $Z_1 + Z_2 + Z_3 \dots + Z_{j+1}$ des scores logarithmiques des j+1 échantillons $f_1, \dots, f_{j+1}$, est calculée.

-a)- si $\log B < Z_1 + \dots + Z_{j+1} < \log A$

le résultat n'est pas significatif, et il faut poursuivre l'enquête, nous sommes dans la région critique.

-b)- si $Z_1 + \dots + Z_{j+1} \geq \log A$

il y a liaison et le pourcentage de recombinaisons est significativement plus petit que x_1 .

-c)- si $Z_1 + \dots + Z_{j+1} \leq \log B$

il n'y a pas liaison, x est significativement plus grand que x_1

Les constantes A et B sont liées à α , probabilité de rejeter H_0 quand H_0 est vraie, et β , probabilité de rejeter H_1 quand H_1 est vraie.

En pratique on utilise deux approximations simples pour la détermination de A et B :

$$A = \frac{1-\beta}{\alpha}$$

$$B = \frac{\beta}{1-\alpha}$$

Il faut enfin fixer la valeur x_1 de x, correspondant à l'hypothèse alternative H_1 . Ce choix se fait arbitrairement selon l'importance des données. En effet, nous avons dit plus haut que l'on appelle efficacité du test, la quantité $1-\beta$. Posons $p(x) = 1-\beta$; c'est une fonction de x qui dépend, de plus, des paramètres x_1 , A et B. Donc, se donner x_1 , A et B, revient à définir de façon précise la fonction $p(x)$. Il est possible de donner l'expression exacte de cette fonction et nous pouvons représenter schématiquement son allure pour différentes valeurs des paramètres : fig. 2 page suivante.

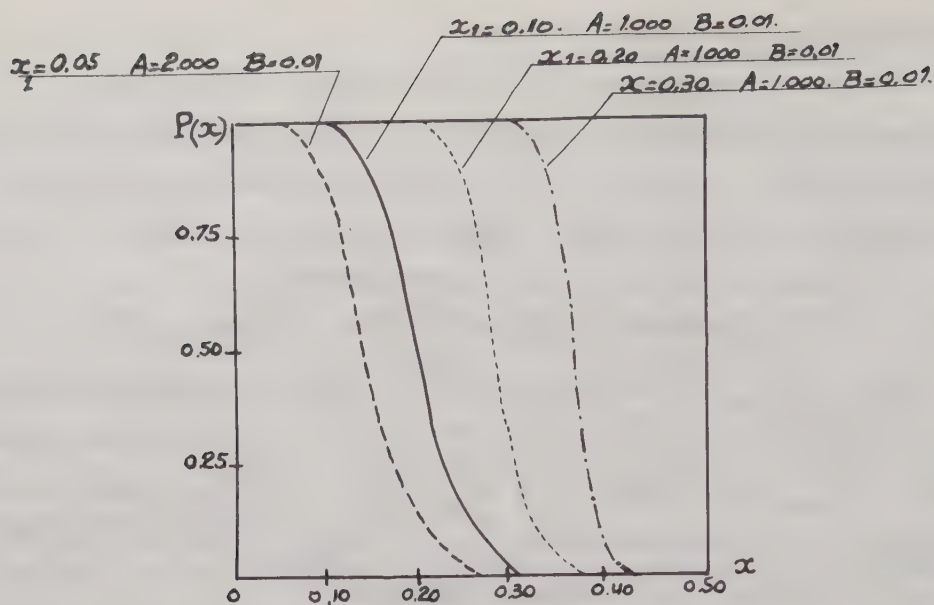


Fig. 2

Variations de $p(x)$ en fonction de x pour différentes valeurs de x . Doubles "test-cross" ou mariages-test dans des fratries de deux enfants.

D'après N.E.MORTON, Sequential Tests for the detection of linkage, Amer. J. of Hum. Genet. 7, 286, 1955.

D'autre part, nous avons vu que le nombre d'opérations nécessaires dépend de la suite des résultats obtenus. C'est donc une variable aléatoire dont il est possible de calculer l'espérance mathématique; ce qui revient à prévoir le nombre approximatif des cas nécessaires pour obtenir un résultat. Cette espérance mathématique peut être calculée de façon rigoureuse; c'est aussi une fonction de x dépendant des paramètres A , B et x_1 . Ses variations sont représentées sur la figure 3, page suivante.

Ces schémas nous montrent tout d'abord que plus x_1 est grand, plus l'efficacité du test est grande. Mais aussi, plus x_1 est grand, plus le nombre moyen d'observations nécessaires est grand. Dans la pratique on dispose d'un certain nombre, plus ou moins important, d'arbres généalogiques. De ce nombre, par lecture des courbes, nous déduisons une valeur optima de x_1 . Ayant fixé x_1 , nous connaissons l'efficacité du test. Si l'on possède un grand nombre d'observations, il sera utile de choisir x_1 grand, ce qui donnera au test une grande efficacité. Si, au contraire, on ne dispose que d'un petit nombre de familles, on sera obligé de prendre x_1 petit et l'efficacité du test diminuera en consé-

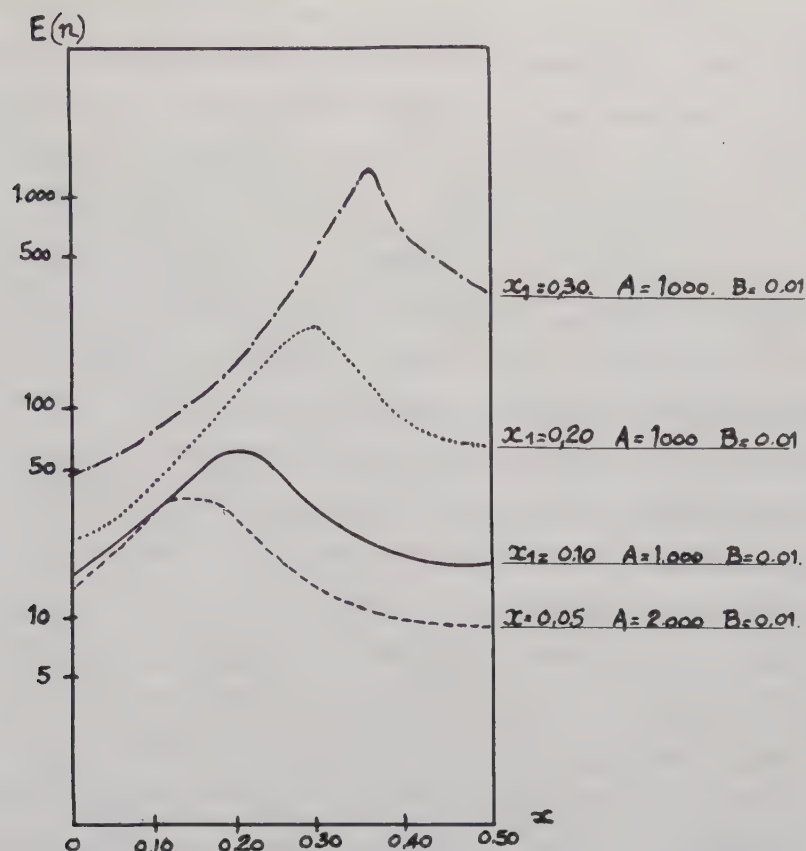


Fig. 3

Espérance mathématique du nombre d'observations nécessaires, $E(n)$, en fonction de x , pour différentes valeurs de x . Double mariage-test, fratrie de deux enfants.

D'après N.E.MORTON, Sequential Tests for the detection of linkage, Amer. J. of Hum. Genet. 7, 287, 1955.

-quence. Cependant, il a été prouvé par MORTON que, dans tous les cas, le nombre d'observations nécessaires pour obtenir un résultat par cette méthode, est inférieur à celui de toutes les autres méthodes non séquentielles. Enfin, autre avantage de cette méthode, il est extrêmement facile de tenir compte de chaque nouvelle information : il suffit d'ajouter au Z déjà calculé, le logarithme du rapport de probabilités $\frac{P_{j+1}(x_1)}{P_{j+1}(1/2)}$ de la nouvelle famille f_{j+1} , soit Z_{j+1} .

- Exploitation pratique des données.

Soit $P(f_i, x, x')$, la probabilité d'apparition de la famille f_i quand le pourcentage de recombinaisons est x chez les femmes et x' chez les hommes. Le score logarithmique de cette famille est défini par :

$$Z^{(i)} = \log_{10} \left\{ \frac{P(f_i | x, x')}{P(f_i | 1/2, 1/2)} \right\}$$

Soient $G, g, G_1, G_2, \dots$ les différents allèles pouvant occuper le locus principal, $T, t, T_1, T_2, \dots$ les différents allèles au locus marqueur. Nous allons maintenant décrire la méthode la plus simple d'utilisation des scores logarithmiques; les familles ou les sujets qui n'apportent que peu ou pas d'information sont exclus du calcul afin de le simplifier. Dans le cas de caractères dominants, il est souhaitable d'analyser la famille tout entière, si l'on s'attend à trouver une liaison. Les scores qui vont être donnés plus loin, s'appliquent aux familles f dans lesquelles les deux parents, et au moins quelques-uns des enfants, ont été examinés; c'est-à-dire que leurs phénotypes sont connus pour le caractère principal et pour le caractère marqueur. (partiellement examiné signifie que l'on connaît seulement le phénotype pour l'un des deux caractères). Il faut, en premier lieu, éliminer les familles suivantes, qui n'apportent pas d'information, et pour lesquelles on ne peut pas calculer de score :

- a)- familles dans lesquelles aucun des deux parents n'est hétérozygote double. Exemple : $Ggtt \times ggTt$
- b)- familles dans lesquelles aucune ségrégation n'est observable : Exemple : $GgTt \times GGTT$, g et t sont récessifs.
- c)- familles dans lesquelles la position des gènes : phase ou répulsion, des parents, est inconnue, et où un seul enfant a été examiné.

Les familles dans lesquelles les génotypes des parents ne sont pas connus avec certitude apportent peu d'information, et le calcul de leur score est pénible. Il vaut donc mieux les omettre.

En général, le score logarithmique "Z" d'une famille, consiste en un score brut au quel s'ajoute une correction "e" ou "d" selon les cas.

Cette correction est nécessaire toutes les fois que le caractère marqueur est récessif et que l'on ne peut connaître avec certitude le génotype des parents au locus marqueur, que par la présence d'un certain phénotype chez l'un au moins des enfants. Ce cas est désigné sous le terme "condition de sélection". En général, les familles qui ne satisfont pas à cette "condition", n'entrent pas dans le calcul du score parce que le type du mariage parental ne peut pas être déterminé avec certitude. Ce point est en effet capital pour l'utilisation des tables de scores logarithmiques. Par exemple, dans un mariage-test tel que $\text{♀ } Tt \times \text{♂ } tt$, où t est récessif, le génotype de la mère

-test tel que : $\text{♀ Tt, x } \text{♂ tt}$, où t est récessif, le génotype de la mère ne peut être connu avec certitude que s'il existe un enfant homozygote récessif. Quelques exceptions à cette règle seront données plus loin.

- Scores pour les types de mariages les plus fréquents.

Pour les valeurs numériques des scores, nous renvoyons le lecteur à l'ouvrage de S.MAYNARD-SMITH, L.S.PENROSE, et C.A.B SMITH : Mathematical tables for research workers in Human Genetics, 1961, Churchill L.T.D, tables 44 à 48, pp. 58 à 63.

Posons $y = 1-x$ et $y' = 1-x'$, pourcentages respectifs de non-recombinaison chez les femmes et chez les hommes.

Dans un test-cross $Gg \times gg$, les phénotypes des enfants seront désignés par $\mathcal{G} = Gg$ et $\mathcal{g} = gg$.

Dans un inter-cross $Gg \times Gg$, ils seront notés $\mathcal{G}^* = GG$ ou Gg et $\mathcal{g}^* = gg$. Le nombre Φ d'enfants d'un phénotype donné qui ont été examinés est désigné par $[\Phi]$. Le nombre total d'enfants dont on tient compte pour le calcul du score Z est s . Les mariages apportant une information sont souvent, nous l'avons vu, des combinaisons de test-cross et d'inter-cross. Les différentes possibilités sont les suivantes ;

-1/- Test-cross - Test-cross : $\text{♀ GgTt} \times \text{♂ gg tt}$

Les enfants sont classés en $\mathcal{G} = Gg$ et $\mathcal{g} = gg$ pour le caractère principal et $\mathcal{T} = Tt$ et $\mathcal{t} = tt$ pour le caractère marqueur. Soit : $a = [\mathcal{GT}] + [\mathcal{gt}]$, c'est-à-dire le nombre d'enfants de génotypes $GgTt$ et $gg tt$ réunis.

$b = [\mathcal{Gt}] + [\mathcal{gt}]$ et $s = a + b$. Soit : $D = [\mathcal{G}]$ et $R = [\mathcal{g}]$.

Le score Z est alors :

$$Z = Z_1 + e_1 \text{ avec :}$$

$$Z_1 = \log_{10} 2^{d-1} \{x^a y^b + x^b y^a\}$$

$$e_1 = \log_{10} \frac{1 - (1/2)^d}{1 - 1/2 x^b y^a - 1/2 x^a y^b}$$

Les valeurs de Z_1 et e_1 sont données dans les tables 44 et 45 de l'ouvrage cité plus haut, pour des familles de 2 à 7 enfants, et des fréquences de recombinaison de 0, 0,05 , 0,1 , 0,2 , 0,3 et 0,4.

-2/- Test-cross - Test-cross : $\text{♀ gg tt} \times \text{♂ Gg Tt}$

L'hétérozygote double est, cette fois, le père, et le score est identique au précédent, à condition de remplacer x par x' et y par y'.

-3/- Test-cross - Inter-cross : $\text{♀ GgTt} \times \text{♂ gg Tt}$, t récessif.

Les enfants sont classés en $G = Gg$, $g = gg$, $T^* = TT$ ou Tt ,

$t^* = tt$. Il y a, dans chaque catégorie,

$$a = [GT^*], b = [Gt^*], c = [gT^*], d = [gt^*]$$

et : $s = a + b + c + d$

avec la condition de sélection : $[t^*] > 0$, $D = [G]$, $R = [g]$

Le score final Z est :

$$Z = Z_2 + d_2$$

avec : $Z_2 = \log_{10} 2^{s-1} / 3^{a+c} [(1+y)^a x^b (1+x)^c y^d + (1+x)^a y^b (1+y)^c x^d]$

$$d_2 = \log_{10} [2^s - (3/2)^s] - \log_{10} [2^s - 1/2(1+y)^b(1+x)^R - 1/2(1+x)^b(1+y)^R]$$

tables 46 et 48, pp. 60 et 63.

-4/- Test-cross - Inter-cross : ♀ $ggTt$ x ♂ $GgTt$, t récessif.

Résultat identique à celui du cas précédent en remplaçant x par x' et y par y' .

-5/- Inter-cross - Test-cross : ♀ $GgTt$ x ♂ $GgTt$, g et t récessifs.

Soient a le nombre d'enfants $[G^*T]$, $b [g^*T]$, $c = [G^*t]$, $d = [g^*t]$ et $s = a + b + c + d$.

avec la condition de sélection : $[t] > 0$, $D = [G^*]$, $R = [g^*]$

le score est :

$$Z = Z_2 + e_2$$

avec : $e_2 = \log_{10} 3^s [1 - 2^{-s}] - \log_{10} [3^s - 1/2(1+y)^D x^R - 1/2(1+x)^D y^R]$

tables 46 et 48.

-6/- Inter-cross - Test-cross : ♀ $GgTt$ x ♂ $GgTt$

Le score est identique à celui du cas précédent en remplaçant x par x' et y par y' .

-7/- Inter-cross - Inter-cross : ♀ $GgTt$ x ♂ $GgTt$, g et t récessifs.

Les enfants sont encore répartis en quatre groupes :

$G^* = GG$ ou Gg , $g^* = gg$, $T^* = TT$ ou Tt , $t^* = tt$ avec :

$$a = [G^*T^*], b = [G^*t^* + g^*T^*], c = [g^*t^*], d = a + b + c$$

condition de sélection : $[t^*] > 0$, $D = [G^*]$, $R = [g^*]$

Le score est :

$$Z = Z_3 + e_3$$

avec :

$$Z_3 = \log_{10} \frac{4^{s-1}}{3^{2a+b}} [(2+y^2)^a x^b (1+y^2)^b y^{2c} + 2(2+xy)^a (1-xy)^b x^c y^c + (2+x^2)^a y^b (1+x)^b x^{2c}]$$

$$e_3 = \log_{10} 3^s [1 - (3/4)^s] - \log_{10} [3^s - 1/4(2+x^2)^D (1-x^2)^R - 1/2(2+xy)^D (1-xy)^R - 1/4 x^{2R} y^{2R} / (2+y^2)^D (1-y^2)^R]$$

Les valeurs de Z_3 et e_3 sont données dans les tables 47 et 48.

-8/- Test-cross - Test-cross : ♀ $GgTt$ x ♂ $ggTt$

Ce type de mariage, nous l'avons dit, ne donne aucune information. Pour les familles ne correspondant pas aux données pour lesquelles le score est fourni par les tables (plus de 7 enfants par exemple) il est possible de calculer ce score à partir des formules suivantes :

$$Z_1(x) = \log_{10} 2^{a-1} [x^a y^b + x^b y^a]$$

$$e_1(x) = \log_{10} [1 - 2^{-a}] - \log_{10} [1 - 1/2 x^a y^a - 1/2 x^b y^b]$$

$$Z_2(x) = \log_{10} 2^{a-1} 3^{-a-c} [(1+y)^a x^b (1+x)^c y^d + (1+x)^a y^b (1+y)^c x^d]$$

$$d_2(x) = \log_{10} [2^a - (3/2)^a] - \log_{10} [2^a - 1/2 (1+y)^a (1+x)^a - 1/2 (1+x)^a (1+y)^a]$$

$$e_2(x) = \log_{10} [3^a (1 - 2^{-a})] - \log_{10} [3^a - 1/2 (1+y)^a x^a - 1/2 (1+x)^a y^a]$$

$$Z_3(x, x') = \log_{10} 4^{a-1} 9^{-a} 3^{-b} [(2+yy')^a (1-yy')^b yy'^c + (2+yx')^a (1-yx')^b \times yx'^c + (2+xy')^a (1-xy')^b xy'^c + (2+xx')^a (1-xx')^b xx'^c]$$

$$e_3(x, x') = \log_{10} 3^a [1 - (3/4)^a] - \log_{10} [3^a - 1/4 (2+xx')^a (1-xx')^a - 1/4 (2+yy')^a (1-yy')^a - 1/4 (2+yx')^a (1-yx')^a - 1/4 (2+xy')^a (1-xy')^a]$$

Dans un certain nombre de cas, aucune correction n'est nécessaire et le score est donné par le Z approprié:

- lorsque le caractère marqueur n'est pas récessif, c'est le cas des groupes sanguins MN, par exemple. Le mariage ♀ GgMM x ♂ GgMN, a pour score $Z_2(x')$.

- lorsque le caractère marqueur est récessif mais qu'il n'y a pas de condition de sélection, c'est-à-dire quand le génotype des parents peut être déterminé par l'examen des phénotypes des ascendants (oncles, tantes, grand-parents). Ainsi, dans le mariage :

♀ GgTt x ♂ gggtt, si la mère Tt, de phénotype T, a un parent récessif, on peut affirmer qu'elle est hétérozygote au locus marqueur sans qu'elle ait nécessairement un enfant récessif, et le score de la famille est $Z_1(x)$.

L'application ou non d'une correction au score ne dépend que du caractère marqueur. Le fait que le génotype d'un parent, au locus principal, est ou non déterminé à partir des enfants n'a aucune importance. Le score $Z(x, x')$ pour l'ensemble des données, est la somme des scores, corrigés si nécessaire, des différentes familles. En général, on suppose $x = x'$ pour simplifier le calcul. Mais, dans certains cas, nous l'avons dit, x peut être différent de x' ; le score $Z(x, x')$ est alors égal à la somme des scores calculés séparément pour des pourcentages de recombinaisons différents dans les deux sexes.

On fixe ensuite la valeur x_1 de x selon l'importance des données; on pose, par exemple, $x = 0,1$ si l'on dispose de quelques familles seulement; $x = 0,2$ si l'on possède un échantillon important et $x=0,3$ si les données sont vraiment très nombreuses. On calcule $Z(x_1)$ en additionnant les scores des familles au fur et à mesure de leur observation. Si, à un moment quelconque, $Z(x_1) > 3$, on peut conclure qu'il y a liaison et que le vrai pourcentage de recombinaisons est inférieur à x_1 . Au contraire, si $Z(x) < -2$, on peut affirmer que le pourcentage de recombinaisons est significativement supérieur à x_1 . Si l'on a $-2 < Z(x_1) < 3$, il faut continuer les observations. 3 et -2 sont les valeurs de $\log A$ et $\log B$ habituellement utilisées.

Les scores Z peuvent également être analysés selon la méthode du maximum de vraisemblance de HALDANE et SMITH (1947). Il suffit de tracer la courbe représentative des variations de $Z(x)$ en fonction de x et d'estimer $\hat{x}$, valeur de x pour laquelle $Z(x)$ est maximum, et

$$I = \left[-d(dx)^2 Z(x) \right]_{x=\hat{x}}$$

c'est-à-dire la pente de la tangente à la courbe au point $x = \hat{x}$.

$\hat{x}$ est l'estimation la plus probable du pourcentage de recombinaisons, avec une erreur standard égale à $1/\sqrt{I}$

On peut encore appliquer la méthode a posteriori ou méthode de BAYES (HALDANE et SMITH, 1947, MORTON, 1955, SMITH, 1959). On construit la courbe représentative de $P(x) = \text{antilog } Z(x)$ et l'on estime la valeur moyenne $\bar{P}$ de $P(x)$ entre $x=0$ et $x=1/2$, c'est-à-dire deux fois l'aire comprise entre la courbe et l'axe des abscisses. La probabilité qu'il y ait une liaison est alors donnée approximativement par :

$$r = \frac{\bar{P}}{\bar{P} + 21 P(1/2)}$$

Pour l'établissement du score dans les cas de mariages autres que ceux cités ci-dessus, nous renvoyons le lecteur à l'ouvrage de S. MAYNARD-SMITH et coll. qui donne une étude détaillée des différents cas possibles.

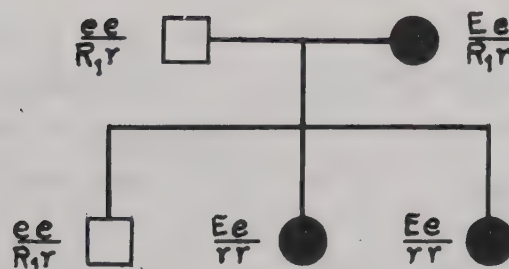
- Exemple d'application de la méthode de N.E.MORTON.

N.E.MORTON a repris par cette méthode l'étude de la liaison entre les loci portant les gènes conditionnant l'elliptocytose et le système Rhésus. Nous représenterons par E le gène de l'elliptocytose, par e son allèle normal. E est dominant, de manifestation régulière, et détermine chez plus de 50% des hématies une déformation elliptique.

Les seuls mariages à considérer sont du type :

$$(Ee) \times (ee)$$

Les mutations, l'illégitimité, etc ... , sont négligeables ou considérées comme telles. Les fréquences des phénotypes Rh sont empruntées à MOURANT, en tenant compte de la nationalité des familles : 7 Anglaises, 2 Suisses, 2 Japonaises, 2 Africaines et 1 Américaine. Le calcul de la probabilité d'apparition d'une famille donnée est conduit comme dans la méthode du rapport de probabilités. Par exemple, pour l'arbre généalogique très simple de CHALMERS et LAWLER (1953) , que nous reproduisons ci-dessous :



on trouve, en représentant les gènes et leurs fréquences par une même lettre :

- s'il y a phase : $P_M = 3/2 \cdot Ee^3 R_1^2 r^2 x^2$

- s'il y a répulsion : $P_R = 3/2 \cdot Ee^3 R_1^2 r^2 (1-x)^2$

d'où :

$$\frac{f(x)}{f(1/2)} = 2 [x^2 + (1-x)^2]$$

et :

$$Z = \log 2 [x^2 + (1-x)^2]$$

Nous rapportons dans le tableau ci-après quelques résultats de MORTON les arbres généalogiques sont classés par ordre chronologique. La lecture de ce tableau montre qu'à partir de la quatrième famille $\sum Z_i > 3$. On peut alors rejeter l'hypothèse $x = 1/2$, c'est-à-dire admettre la liaison avec un risque d'erreur de l'ordre de $1/1000$.

Numéro d'identification des pedigrees	Valeurs de α			
	0,05	0,10	0,20	0,30
1	0,2577	0,2148	0,1335	0,0645
2	— 4,2320	— 2,1820	— 0,5401	0,0178
3	4,6821	4,2360	3,2641	2,1650
4	2,3213	2,3170	1,9531	1,3763
$z_1 + z_2 + z_3 + z_4$. . .	3,0291	5,1858	4,8106	3,6236
Total pour les 14 pedigrees	2,70	6,14	7,41	5,57

Pour l'ensemble des familles et pour $\alpha = 0,20$, $\sum z_i = 7,41$.

Prenons A tel que $\log A = 7,41$; $\alpha = \frac{1-\beta}{A} < \frac{1}{A}$, c'est-à-dire, dans notre cas, $\alpha < \frac{1}{\text{antilog } 7,41} = 3,9 \times 10^{-3}$. La liaison est donc hautement significative.

Nous pouvons maintenant, pour terminer, essayer de préciser les caractéristiques et principaux avantages et inconvénients de ces différentes méthodes.

- a)- la méthode des paires de germains de L.S.PENROSE, très utilisée en raison de sa grande simplicité, s'applique aux cas où les génotypes des parents sont mal connus ou inconnus, et aussi, lorsque le déterminisme génétique des caractères étudiés n'est pas connu avec précision. Elle ne nécessite que la connaissance d'une génération sous la forme d'un grand nombre de paires de germains; mais ces paires doivent être indépendantes en probabilité, et la partition d'une famille de taille $n > 2$ en $\frac{n(n-1)}{2}$ paires, comme cela a souvent été fait à tort, donne des résultats inexacts. C.A.B.SMITH a proposé des corrections pour compenser la non-indépendance des paires, mais elles suppriment le principal avantage de la méthode, sa grande simplicité d'utilisation.
- b)- La méthode des scores-U de R.A.FISHER s'applique à des fratries

de deux germains ou plus, lorsque le génotype de l'un des parents au moins est connu. Grâce à D.J.FINNEY, elle peut être utilisée pour tous les types de mariages mais son emploi est long et complexe et C.A.B.SMITH a résumé ainsi ses inconvénients :

- bien que les valeurs U soient très faciles à utiliser quand le génotype des deux parents est complètement connu (sauf en ce qui concerne la position en phase ou en répulsion), le calcul de la variance peut être impossible s'il y a des inconnues. Lorsque les échantillons sont grands, cette difficulté peut être surmontée par l'utilisation d'une approximation simple.
 - la quantité d'information donnée par ce test est une fonction de la valeur alternative x_1 ; cette information est grande surtout pour des valeurs de x_1 voisines de $1/2$, donc dans les cas de liaison lâche. Or ceux-ci sont difficiles à déceler et comportent de grands risques d'erreur. Un test efficace pour des valeurs moyennes de x_1 serait plus intéressant.
 - l'information peut être très augmentée par l'utilisation d'observations portant sur trois générations ou plus. Il n'est pas possible d'obtenir cette information supplémentaire par l'utilisation des scores U.
 - enfin, cette méthode, comme la précédente, permet de reconnaître l'existence d'une liaison entre deux loci, mais ne donne pas d'estimation précise ni de limites de confiance du pourcentage de recombinaisons. Pour toutes ces raisons, nous n'avons pas donné d'exposé de la méthode de FISHER et FINNEY, méthode qui, actuellement est moins utilisée que celles dont nous venons de parler dans ce deuxième chapitre.
- c)- la méthode séquentielle de H.E.MORTON, la dernière en date, est nettement plus simple et plus efficace que la plupart des méthodes utilisées auparavant, et d'une utilisation pratique aisée. Cependant son emploi soulève d'importantes questions de principe (C.A.B.SMITH) On peut se demander en effet, si l'utilisation d'une méthode séquentielle de WALD est vraiment adaptée aux études de liaison, et si elle présente un quelconque avantage sur les méthodes non séquentielles. Pour C.A.B.SMITH, le théorème de BAYES est d'une application presque aussi simple et donnerait des résultats plus satisfaisants des deux points de vue, théorique et pratique. En fait, les réponses données par les deux méthodes sont pratiquement identiques, et la discussion ne se place guère que sur un plan théorique : a-t-on ou non le droit

pour des études de liaisons en Génétique Humaine, d'adopter une règle d'arrêt séquentielle?

En ce qui concerne l'application du théorème de BAYES, qui théoriquement paraît idéale, l'obstacle majeur, en général impossible à surmonter dans la pratique, est l'absence quasi-totale de données concernant les probabilités "initiales" ou a priori. Cette méthode est pourtant la seule à donner directement une réponse en terme de probabilité, sans avoir à faire intervenir une limite de confiance. Le chercheur qui utilise le théorème de BAYES peut arrêter son échantillon quand il le désire, pour quelque motif que ce soit, et, à condition de faire porter son calcul sur l'ensemble des données réunies, il peut fournir un résultat parfaitement correct exprimé sous la forme de la probabilité de l'existence de la liaison. Ceci est, évidemment, un très grand avantage. Le théorème de BAYES permet, de surcroît, d'obtenir les probabilités des différentes valeurs de x comprises entre 0 et 0,5. Il permet donc de combiner en un seul calcul, la mise en évidence et l'estimation de l'intensité de la liaison. Grâce à l'aide considérable apportée par les tables de MORTON, ceci ne nécessite, finalement, qu'assez peu de calculs. Cependant il faut insister sur le fait que l'application de cette méthode repose sur la validité de l'hypothèse initiale de normalité de la distribution de x . Cette hypothèse, bien que très plausible, peut n'être qu'une approximation, et, par conséquent, les probabilités calculées ne sont, elles aussi, que des approximations. Ceci, dans l'état actuel de nos connaissances, paraît un obstacle inévitable.

Il n'existe donc pas, à l'heure actuelle, de méthode réunissant tous les suffrages par sa simplicité, son exactitude quelle que soit l'imprécision des données. Mais il existe au moins trois bons instruments statistiques capables de donner une approximation suffisante, et assez satisfaisante.

- CONCLUSIONS -

Au cours de ce travail, nous avons essayé de donner une vue d'ensemble des problèmes que pose l'établissement de la carte chromosomique en Génétique Humaine, et des différents moyens de les résoudre.

Nous avons traduit le terme anglo-saxon de "linkage" par "liaison entre loci" pour insister sur le fait qu'il s'agit, non pas d'une liaison "génétique" comme le disent, et le pensent encore, de trop nombreux auteurs, mais de la liaison entre divers "emplacements" ou loci, situés sur un même chromosome.

Nous avons ensuite indiqué les principales difficultés du sujet. D'ordre pratique d'abord :

En premier lieu, il ne faut pas confondre liaison entre loci et association entre caractères, celle-ci n'étant qu'une forme très particulière de celle-là, la seule constatation d'une association entre caractères ne suffisant pas pour évoquer la proximité, sur un même chromosome, des gènes responsables.

Nous avons souligné qu'actuellement, une des difficultés les plus aiguës, est le manque de gènes dits "marqueurs", qu'il faudrait pouvoir connaître en nombre tel que tous les chromosomes humains soient marqués de 20 en 20 centimorgans, ce dont nous sommes malheureusement encore très loin.

Nous avons aussi insisté sur la pénurie des familles "riches d'information", c'est-à-dire des familles issues de l'union d'un ou une hétérozygote et du partenaire correspondant homozygote pour un gène récessif, c'est-à-dire d'un test-cross, mariage le plus favorable pour ces études; puisque, en Génétique Humaine, nous ne pouvons qu'observer des familles déjà constituées et leurs descendance, et non intervenir dans la formation des couples.

Ensuite, nous avons aussi attiré l'attention sur les difficultés beaucoup plus complexes, d'ordre mathématique et statistique.

Enfin, nous avons passé en revue les différentes méthodes qui ont été, ou sont actuellement utilisées dans ces études :

méthodes de FISHER, de PENROSE, pour insister plus particulièrement sur la méthode de N.E.MORTON, dont l'application est la plus aisée et paraît la plus conforme aux données du problème. Le principe de cette dernière méthode repose sur l'analyse séquentielle. Elle consiste à calculer un certain score "Z" pour une ou plusieurs familles, et à le comparer à deux nombres A et B, l'un supérieur à 1, l'autre inférieur à 1, définis par les expressions : $A = \frac{1-\beta}{\alpha}$, $B = \frac{\beta}{1-\alpha}$ α et β étant les erreurs de 1^{re} et 2^{de} espèces.

Si Z est supérieur à A, on peut penser, avec un certain pourcentage de chances qui est fonction des valeurs de α et β choisies a priori au départ, qu'il y a liaison entre les loci envisagés. Si Z est inférieur à B, on peut, avec le même pourcentage de sécurité, rejeter l'hypothèse de liaison. Enfin, si Z est compris entre A et B, il faut continuer l'étude, c'est-à-dire rechercher de nouvelles familles susceptibles d'apporter une information et calculer le nouveau score Z de cet ensemble de familles. Il est démontré que l'on finit toujours par obtenir un score Z, soit supérieur à 1, soit inférieur à 1. On peut donc toujours conclure, soit à l'existence d'une liaison, soit à l'indépendance des loci étudiés. Divers exemples d'application de ces méthodes sont donnés.

Nous ferons maintenant très rapidement le point des quelques liaisons actuellement mises en évidence avec une quasi-certitude.

-I- Liaisons autosomiques.

- Onycharthrose et groupe sanguin ABO.

S.D.LAWLER et J.H.RENWICK, plus récemment (1965) J.H.RENWICK et J.SCHULZE, ont donné une estimation de x qui varie selon le sexe, phénomène déjà observé dans certaines espèces animales

$$x_{\text{f}} = 0,146 \quad (\text{SE} = 0,032) \quad x_{\text{m}} = 0,084 \quad (\text{SE} = 0,023)$$

Il semble actuellement que le sexe homogamétique montre plus de recombinaisons pour cette paire de loci que le sexe hétérogamétique (règle de HALDANE, 1922).

- Elliptocytose et Rh.

C.A.CLARKE, H.B.GOODALL, D.W.W.HENDRY, S.D.LAWLER, S.A.STEPHEN, R.A.MARSHALL et N.E.MORTON, ensemble ou séparément ont donné en dernière estimation : $x = 0,20$

- Groupes Lutheran et Secréteur.

R.CEPPELLINI et M.SINISCALCO, E.KAKLAMANIS, E.J.MOLBOROW et L.E.GLYNN, P.J.L.COOK enfin, ont mis en évidence cette liaison qui, elle aussi, semble présenter une intensité différente dans les deux sexes.

$x =$

Enfin, tout récemment, puisque ces travaux datent de 1965, une nouvelle liaison autosomique semble avoir été mise en évidence entre les loci de "Xeroderma Pigmentosum" et, du groupe ABO, (ce qui porterait à trois le nombre des loci du "groupe de linkage" ABO), par H.EL HEFNANI, S.MAYNARD-SMITH et L.S.PENROSE.

$x = 0,179$

Ceci porte donc à quatre le nombre des liaisons autosomiques actuellement reconnues. On mesure aisément l'ampleur du travail à poursuivre si l'on sait que le nombre des loci humains est estimé à environ 6×10^4

-II- Loci portés par le chromosome X.

En ce qui concerne le chromosome sexuel, le problème est nettement plus simple puisqu'il s'agit seulement de placer des loci les uns par rapport aux autres et par rapport au centromère. Ceci explique que l'on puisse commencer à parler de l'ébauche de la carte factorielle de ce chromosome. Nous résumerons simplement les résultats obtenus sur un petit schéma, dont il faut bien préciser que, dans l'état actuel de nos connaissances, il est impossible de dire qu'il représente autre chose que des hypothèses qui peuvent très bien ne pas être en accord parfait avec la carte chromosomique réelle du X. Nous citerons les nombreux auteurs qui aboutirent à ce résultat dans la partie de la bibliographie réservée aux résultats.

<u>Nom de la liaison</u>	<u>valeur de x</u>
Xg - G6PD	27 cmorg.
Cécité aux couleurs - Hémophilie A.	12 cmorg.
Cécité aux couleurs - G6PD.	5 cmorg.

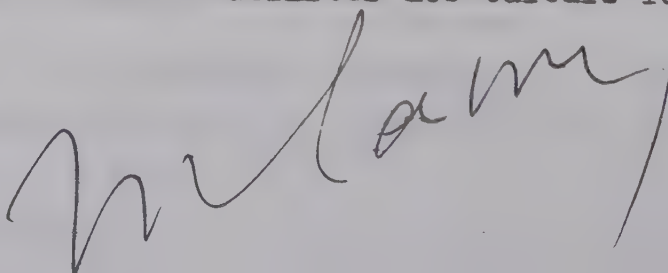
Nous ajouterons que les techniques de la Génétique Cytologique semblent maintenant ouvrir une nouvelle voie de recherches, très féconde, pour ces études et il est permis de penser que l'étude des ~~aberrations~~ chromosomiques permettra de compléter utilement les procédés purement mathématiques. C'est ainsi que l'étude clinique et métabolique des Mongoliens, porteurs d'une trisomie 21, a permis de penser que les loci contrôlant l'activité de la galactokinase ou de la phosphohexokinase par exemple, sont situés sur ce chromosome; en effet, ces sujets ont un taux enzymatique correspondant à 1,5 fois le taux observé chez les sujets normaux.

Ceci permet de constater aisément que ces études qui paraissent au premier abord très mathématiques et théoriques ont une utilité directement pratique; l'information fournie par les études de liaisons entre loci a, en effet, une valeur certaine en tant qu'instrument de recherche. Elle a, en particulier, été utilisée pour démontrer qu'une seule manifestation phénotypique, telle que l'elliptocytose, peut être due à une mutation à un locus dans une famille, et à une mutation à un autre locus dans une autre famille (MORTON, N.E., 1956, LAWLER et coll 1954). Ainsi, chez l'Homme, où il est impossible de pratiquer un test direct d'allélisme, il est particulièrement utile de posséder une évidence indirecte, sorte de démonstration par l'absurde, de non allélisme, grâce à des relations différentes de liaison à un troisième locus, rhésus dans l'exemple choisi.

Inversement, quoique moins nettement dans ce sens, toute probabilité que deux gènes soient allèles se trouve augmentée si on les trouve équidistants d'un même locus marqueur. Ceci pourrait, théoriquement, s'appliquer à l'étude d'une série d'allèles normaux ou isoallèles, qui ne peuvent être différenciés l'un de l'autre que par leur manifestation phénotypique lorsqu'ils sont en combinaison hétérozygote avec un allèle mutant au même locus.

Pour terminer, nous dirons tout l'espoir que donne aux chercheurs en ce domaine, l'emploi des calculatrices électroniques

, qui semblent devoir singulièrement faciliter des calculs longs et fastidieux jusqu'alors.



BIBLIOGRAPHIE

- BAILEY, NTJ. 1- The use of the product formula for the estimation of linkage in intercrosses when differential viability is present. *Heredity* 3, 220-225.
- 2- A method of allowing for differential viability in estimating linkage from back-cross matings in coupling only or in repulsion only. *Heredity* 3, 225-228.
- 3- The influence of partial manifestation on the detection of linkage. *Heredity*, 4, 327-336, 1950.
- 4- On simplifying the use of Fisher's u-statistics in the detection of linkage in man. *Ann. Eugen. Lond.* 16, 26-32, 1951.
- 5- Introduction to the mathematical theory of Genetic linkage. Oxford Clarendon Press, 1961.
- BODMER, WF . 1- Multiplicative effects and the logarithmic transformation in the analysis of balanced multi-point linkage tests. *Heredity*, 13, 157-164. 1959.
- 2- The analogy between factorial experimentation and balanced multipoint linkage tests. *Heredity*, 13, 145-156, 1959.
- 3- Factorial analysis of balanced four and higher-point linkage tests. *Heredity*, 13, 487-498. 1959.
- EDWARDS, AWF. 1- Number of mating types required for balance in multi-point linkage programmes. *Nature*, 181, 503-504, 1958.
- FINNEY, DJ. 1- The detection of linkage. *Ann. Eugen. Lond.* 10, 171-214, 1940.
- 2- The detection of linkage. II. Further mating types; scoring of BOYD's data. *Ibid*, 11, 10-30, 1941.
- 3- The detection of linkage. III. Incomplete parental testing. *Ibid*, 11, 115-135, 1941.
- 4- The detection of linkage. IV. Lack of parental records and the use of empirical estimates of information. *J. Hered.* 33, 157-160, 1942.
- 5- The detection of linkage. V. Supplementary tables. *Ann. Eugen. Lond.* 11, 224-232, 1942.
- 6- The detection of linkage. VI. The loss of information from incompleteness of parental records. *Ibid*. 233-242, 1942.
- 7- The detection of linkage. VII. Combination of data from matings of known and unknown phase. *Ann. Eugen. Lond.* 12, 31-43, 1943.
- 8- The estimation of the frequency of recombinations. Matings of known phase. *J. Genet.* 49, 159-176, 1949.

FISHER, RA. 1- The effects of methods of ascertainment upon the estimation of frequencies. Ann. Eugen. Lond, 6, 13-25, 1934.

2- The amount of information supplied by records of families as a function of the linkage in the population sampled. Ann. Eugen. Lond, 6, 66-70, 1934.

3- The use of simultaneous information estimations in the evaluation of linkage. Ibid, 71-76.

4- The detection of linkage with "dominant abnormalities". Ibid, 187-201, 1935.

5- Tests of significance applied to HALDANE's data on partial sex-linkage. Ann. Eugen. Lond. 7, 87-104, 1936.

6- The precision of the product formula in the estimation of linkage. Ibid, 50-54.

7- The theory of linkage in polysomic inheritance. Phil. Trans. B, 233, 55-87, 1947.

8- A quantitative theory of genetic recombination and chiasma formation. Biometrika, 4, 1-13, 1948.

9- A combinatorial formulation of multiple linkage tests. Nature, 167, 520, 1951.

10- Statistical tables for biological, agricult. and medical research. Lond. Oliver Boyd, 1953.

GEIRINGER, H. 1- On the probability theory of linkage in Mendelian heredity. Ann. Math. Statist. 15, 25-57, 1944.

2- Further remarks on linkage in Mendelian heredity. Ibid, 16, 390-393

3- On the definition of distance in the theory of the gene. Ibid. 16, 393-398, 1945.

HALDANE, JBS. 1- The combination of linkage values and the calculation of distance between the loci of linked factors. J. Genet, 8, 299-309, 1919.

2- Methods for the detection of autosomal linkage in man. Ann. Eugen. Lond, 6, 26-65, 1934.

3- The cumulants of the distribution of FISHER's u and u^4 scores used in the detection and estimation of linkage in man. Ibid. 13, 122-3

HOGBEN, LT. 1- The detection of linkage in human families. Proc. Roy. Soc. B, 114, 340-363, 1934.

2- An introduction to mathematical genetics. New York, Norton, 1946.

IMMER, FR. 1- Formulae and tables for calculating linkage intensities. Genetics, 15, 81-98, 1930.

2- Calculating linkage intensities from F data. Ibid, 19, 119-136, 1934

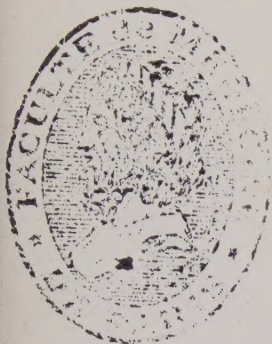
- KOSAMBI, DD. The estimation of map distance from recombination values. Ann. Eugen. Lond. 12, 172-175, 1944.
- MALECOT, G. Mathématiques de l'Hérédité. Masson, 1948.
- MATHER, K. 1- The measurement of linkage in heredity. 2^e ed. London Methuen, 1951.
2- Statistical analysis in biology. Methuen et Co. London, 1951.
- MORTON, NE. 1- Sequential tests for the detection of linkage. Amer. J. Hum. Genet, 7, 277-318, 1955.
2- The detection and estimation of linkage between the genes for elliptocytosis and the Rh blood type. Ibid, 8, 80-96, 1956.
3- Genetic tests under multiple ascertainment. Ibid, 11, 1-16, 1959.
- OAKLAND, GB. The statistical analysis of genetic linkage data. Acta genetica et statistica medica, Basel, 10, n°1-3, 191-236, 1960.
- OWEN, ARG. The analysis of multiple linkage data. Heredity, 7, 247-64, 1953.
- PENROSE, LS. The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. Ann. Eugen. Lond, 6, 133-138, 1935.
2- Genetic linkage in graded human characters. Ibid, 8, 233-238, 1938.
3- A further note on the sib-pair linkage method. Ibid, 13, 25-29, 1946.
4- Data for the study of linkage in man: phenylketonuria and the ABO and MN loci. Ibid, 16, 241-248, 1950.
5- Data for the study of linkage in man: red-hair and ABO locus. Ibid, 15, 243-247, 1950.
6- The general purpose sib-pair linkage test. Ibid, 18, 120-124, 1953.
7- Measurements of likeness in relatives of trisomics. Ann. Hum. Genet 27/2, 183-187, 1963.
- RENWICK, JH. 1- Close linkage between ABO and nail patella loci. Ann. Hum. Genet. 19, 312, 1955.
2- Linkage in human Genetics. The eugenics review, vol 48, n°5, Oct. 1956.
3- Blood groups and genetic linkage. British Med. Bull. vol 15, N°2, 145-149, 1959.
4- Elucidation of gene order. Recent advances of Human Genetics, LS, PENROSE, 1961.
5- A computer programme for the processing of linkage data from large pedigrees. Proc. II, Intern. Conf. Hum. Genet. Rome, 1961.

- RIFE, DC. Populations of hybrid origin as source material for the detection of linkage. Amer. J; Hum. Genet. 6,1, 26-33, 1954.
 - ROBERTS, JAF. Genetic linkage in man, with particular reference to the usefulness of very small bodies of data. Quart. J; Med. 14, 27-33, 1945.
 - RUFFIE, J. Les groupes sanguins chez l'homme. Masson, 1955
2- en coll. avec HURON, Les methodes en Génétique générale et en Génétique Humaine. Masson, 1959.
 - SIMPSON, HR. The estimation of linkage on an electronic computer. ANN, Hum. Genet. Lond. 22, 356-351, 1958.
 - SMITH, CAB. 1- The detection of linkage in Human Genetics. The JR Statist. Soc. B, 15, 153-192, 1953.
2- The separation of the sexes of parents in the detection of linkage in man. Ann. Eugen. Lond. 18, 278-301, 1954.
3- A test for segregation ratios in family data. Ibid. 20, 257-265, 1956.
4- Some comments on the statistical methods used in linkage investigations. Amer. J; Hum. Genet. 11, n°4, 289-304, 1959.
5- Biomathematics. Hafner, New-York, 1954.
6- Testing for heterogeneity of recombination fraction values in human genetics. ANN. Hum. Genet. 27/2, 175-182, 1963.
 - WU, TT. A note on genetic mapping. A model is introduced to make it easy to calculate the recombination rate as a function of the map distance. Haldane's (1918) experimental result is used for comparison. J. Theor. Biol. 713, 430-434, 1964.
-

FACULTE DE MEDECINE DE PARIS

ANNEE 1966

N° 64762



THESE

pour le

DOCTORAT EN MEDECINE
(Diplôme d'Etat)

par

VU:
LE DOYEN

Edith LEDINOT-DUBREIL

Ancienne Externe des Hôpitaux de PARIS

Née le 22 Septembre 1936 à PARIS XVI°

Pour le Doyen
L'Assesseur,

Présentée et soutenue publiquement le

1966 .

Les liaisons entre loci en Génétique Humaine

(Problèmes théoriques)

Président : M. Maurice LAMY, Professeur.

VU
et PERMIS D'IMPRIMER
LE RECTEUR
DE L'ACADÉMIE DE PARIS

